

**Thunderstone Search Appliance
WWW Site Indexer Version 6.2.4**

Thunderstone Software

February 21, 2007

Contents

1	Overview	13
1.1	Features	13
1.2	Technical Support	14
2	Installation	15
2.1	How to unpack and install the Search Appliance	15
2.2	Customizing the Search Appliance's Appearance	15
3	Operation	17
3.1	Running the Administrative Interface	17
3.2	First Time Run	17
3.3	Administrative Interface Overview	19
3.3.1	Entry	19
3.3.2	Basic Walk Settings	20
3.3.3	All Walk Settings	20
3.3.4	Search Settings	21
3.3.5	Best Bet Groups	21
3.3.6	List/Edit URLs	21
3.3.7	Walk Status	22
3.3.8	Query Log	23
3.3.9	Test Search	24
3.3.10	Live Search	24
3.3.11	Profiles	24
3.3.12	Accounts	24

3.3.13	User Groups	25
3.3.14	Access Control	26
3.3.15	Maintenance	26
3.3.16	Documentation	26
3.3.17	Appliance Home	26
3.3.18	Logout	26
3.4	Basic Walk Settings	26
3.4.1	Walk Summary	26
3.4.2	Notes	27
3.4.3	Base URL	27
3.4.4	Enterprise	27
3.4.5	Robots	28
3.4.6	Extensions	28
3.4.7	Exclusions	29
3.4.8	Crawl Delay	29
3.4.9	Parallelism	29
3.4.10	Verbosity	29
3.4.11	Rewalk Type	29
3.4.12	Rewalk Schedule	31
3.4.13	Action Buttons	31
3.5	Advanced Walk Settings	32
3.5.1	Watch URL	32
3.5.2	Notify	32
3.5.3	Attach Logs	32
3.5.4	Categories	32
3.5.5	DBWalker	33
3.5.6	URL File	33
3.5.7	URL URL	33
3.5.8	Single Page	33
3.5.9	Page File	33

3.5.10	Page URL	34
3.5.11	Strip Queries	34
3.5.12	Ignore Case	34
3.5.13	Extra Domains	34
3.5.14	Extra Networks	35
3.5.15	Extra URLs REX	35
3.5.16	Exclusion REX	35
3.5.17	Exclusion Prefix	36
3.5.18	Exclude by Field	36
3.5.19	Additional Fields	36
3.5.20	Data from Field	37
3.5.21	Required REX	37
3.5.22	Required Prefix	37
3.5.23	Max Page Size	37
3.5.24	Max Pages	37
3.5.25	Max Bytes	38
3.5.26	Max Depth	38
3.5.27	Page Timeout	38
3.5.28	Meta Tags	38
3.5.29	Standard Meta	38
3.5.30	All Meta	38
3.5.31	Storage Charset	39
3.5.32	Source Default Charset	39
3.5.33	XML UTF-8	39
3.5.34	Keep HTML	39
3.5.35	Keep Links	40
3.5.36	Remove Common	40
3.5.37	Ignore Tags	40
3.5.38	Keep Tags	40
3.5.39	Ignore Characters	41

3.5.40	Plugin Split	41
3.5.41	Word Definition	41
3.5.42	Index Fields	42
3.5.43	Primer URL	42
3.5.44	Login Info	42
3.5.45	Proxy	43
3.5.46	Proxy Login Info	43
3.5.47	Cookie Source Path	43
3.5.48	Off-Site Pages	43
3.5.49	Stay Under	43
3.5.50	Prevent Duplicates	44
3.5.51	Duplicate Check Fields	44
3.5.52	All Extensions	44
3.5.53	Store Refs	44
3.5.54	Inline Iframes	45
3.5.55	Max Frames	45
3.5.56	Execute JavaScript	45
3.5.57	Fetch JavaScript	45
3.5.58	JavaScript String Links	45
3.5.59	Debug JavaScript	46
3.5.60	Protocols	46
3.5.61	Embedded Security	46
3.5.62	Entropy Source	46
3.5.63	Max Redirects	46
3.5.64	Index Name	46
3.5.65	DNS Mode	47
3.5.66	User Agent	47
3.5.67	Mime Types	47
3.5.68	Respect Expires Header	47
3.5.69	Default Refresh Time	47

3.5.70	Minimum Refresh Time	48
3.5.71	Maximum Refresh Time	48
3.5.72	Maximum Process Size	48
3.5.73	Replication Settings	48
3.6	Search Settings	48
3.6.1	Notes	48
3.6.2	Query Logging	49
3.6.3	Rotate Schedule	49
3.6.4	Email	49
3.6.5	Result Order	49
3.6.6	Results Style	49
3.6.7	XSL File	50
3.6.8	Abstract Style	50
3.6.9	Abstract Length	50
3.6.10	Results per Page	50
3.6.11	Results per Site	50
3.6.12	Allow site: syntax	50
3.6.13	Results Width	51
3.6.14	Box Color	51
3.6.15	Display Thunderstone logo on results	51
3.6.16	Show Advanced	51
3.6.17	Font	51
3.6.18	Display Charset	51
3.6.19	Top HTML and Bottom HTML	52
3.6.20	Enable Sherlock	52
3.6.21	Apply Appearance and Revert Appearance	52
3.6.22	Top Best Bet Title	53
3.6.23	Right Best Bet Title	53
3.6.24	Top Best Bet Group	53
3.6.25	Right Best Bet Group	53

3.6.26	Top Best Bet Box Color	53
3.6.27	Right Best Bet Box Color	53
3.6.28	Top Best Bet Border Style	54
3.6.29	Right Best Bet Border Style	54
3.6.30	Right Best Bet Box Width	54
3.6.31	Authorization Method	54
3.6.32	Login Cookies	54
3.6.33	Login URL	55
3.6.34	Basic/NTLM/file Cookie Type	55
3.6.35	Login Verification URL	55
3.6.36	Unauthorized Result Query	56
3.6.37	Max Docs to Auth-Check	56
3.6.38	Successful Auth Result Limit	56
3.6.39	Total Auth Timeout	57
3.6.40	Debug Results Authorization	57
3.6.41	Enable Spell Check	57
3.6.42	Suggest Time Limit	57
3.6.43	Number of Suggestions	57
3.6.44	Synonyms	58
3.6.45	Main Thesaurus	58
3.6.46	Secondary Thesaurus	58
3.6.47	Allow the @ Operator	58
3.6.48	Allow Linear	59
3.6.49	Allow NOT Logic	59
3.6.50	Allow Post-Processing	59
3.6.51	Allow Wildcards	59
3.6.52	Allow WITHIN Operators	59
3.6.53	Resolve Phrase Noise Words	60
3.6.54	Keep Noise Words	60
3.6.55	Search Timeout	60

3.6.56	Fast Result Counts	60
3.6.57	Proximity	60
3.6.58	Word Forms	61
3.6.59	Word Ordering	61
3.6.60	Word Proximity	61
3.6.61	Database Frequency	61
3.6.62	Document Frequency	62
3.6.63	Position in Text	62
3.6.64	Ranked Rows	62
3.6.65	XML Export Variables	62
3.6.66	File Url Format	62
3.6.67	Visible	63
3.7	Results Authorization	63
3.7.1	Results Authorization Crawl Settings	64
3.7.2	Results Authorization Search Settings	64
3.8	Meta Search - Search multiple profiles as one	64
3.8.1	Profile Creation	64
3.8.2	Walk Settings	64
3.8.3	Search Settings	65
3.9	Access Control	65
3.9.1	User Groups	65
3.9.2	Object hierarchy	66
3.9.3	Access Control Lists	66
3.9.4	Determining Effective Rights	67
3.9.5	Required Rights for Admin Actions	67
3.10	Running the Search Interface	69
3.11	Maintenance	70
3.11.1	Information	70
3.11.2	Install/Upgrade	71
3.11.3	Logs	71

3.11.4	Search Appliance Settings	72
3.11.5	Appliance system access	78
4	Procedures and Examples	83
4.1	Searching your Index	83
4.2	Similarity Searching	84
4.3	Using the Thesaurus Feature	85
4.4	Getting Software Updates	86
4.5	Page Exclusion, Robots.txt, and Meta-robots	86
4.6	Indexing Other Sites	88
4.7	Indexing Individual Pages	88
4.8	Reindexing on a Schedule	89
4.9	Checking for Web Server Errors	89
4.10	Removing Pages from the Database	89
4.11	Erasing the Entire Database	89
4.12	Using Multiple Databases	89
4.13	Integrating the Search Appliance with your Site	89
4.13.1	Static Host	90
4.13.2	Dynamic Host and HTML	91
4.13.3	Dynamic Host and XML	91
4.14	Using Best Bets	94
4.15	Using Access Control	94
4.15.1	Initial Lockdown	94
4.15.2	Example: User with Complete Control on One Profile	95
4.15.3	Example: User with Look and Feel Control on All Profiles	95
4.16	Indexing File Servers	95
4.17	Replication	96
4.17.1	Replication Overview	96
4.17.2	Procedure	96
4.17.3	DataLoad API	97
4.18	Additional Fields	102

4.18.1	Overview	102
4.18.2	Populating	102
4.18.3	Sorting	102
4.18.4	Searching	102
4.19	DBWalker	103
4.19.1	Overview	103
4.19.2	Configuration Overview	103
4.19.3	DBWalker Output Overview	104
4.19.4	DBWalker Authentication Overview	104
4.19.5	Obtaining DBWalker	104
4.19.6	Managing DBWalker	105
4.19.7	DBWalker Global Options	105
4.19.8	Managing DBWalker Configurations	106
4.19.9	Managing DBWalker Stylesheets	109
4.19.10	Adding Configurations to Profiles	109
4.20	Thunderstone Proxy Module	109
4.20.1	Overview	109
4.20.2	Requirements	110
4.20.3	Installing the Proxy Module	110
4.20.4	Configuring the Search Appliance	111
4.20.5	Manually Configuring the Proxy Module	113
4.20.6	Manual Installation Steps	113
5	Reference	117
5.1	Database and File Usage	117
5.2	Walk Database Tables and Fields	118
5.3	Options Table Fields	120
5.4	Customizing the Search	121
5.5	Customizing the Walker	121
5.6	Third-Party Software	121
5.6.1	Antiword	121

5.6.2	Aspell	121
5.6.3	Catdoc xls2csv	121
5.6.4	Cole library	122
5.6.5	iconv	122
5.6.6	JDBC drivers	122
5.6.7	ppt2html, msg2html	127
5.6.8	SSL/HTTPS plugin	127
5.6.9	unrar	130
5.6.10	unzip	131
5.6.11	zlib	132
5.6.12	SpiderMonkey (JavaScript-C) Engine	132
5.6.13	PDF/anytotx plugin	132
5.6.14	thttpd - throttling HTTP server	133
5.6.15	RedHat Linux	133
5.6.16	Webmin	133
5.6.17	Java	134
5.6.18	OpenSSL RPM	139
5.6.19	RAID utilities	139
5.6.20	GNU General Public License	140
5.6.21	GNU Lesser General Public License	147
5.6.22	GNU Library General Public License	157
5.6.23	Netscape Public License	166
5.7	XML Elements in Search Results	175
6	Search Interface Help	177
6.1	Forming a Query	177
6.1.1	Query Rules of Thumb	177
6.1.2	Overview of Query Abilities	178
6.1.3	Controlling Proximity	178
6.1.4	Ranking Factors	178
6.1.5	Keywords Phrases and Wild-cards	178

6.1.6	Applying Search Logic	179
6.1.7	Natural Language Query	180
6.1.8	Using the Special Pattern Matchers	180
6.1.9	Invoking Thesaurus Expansion	181
6.2	Using Word Forms	181
6.3	Controlling Proximity	182
6.4	Interpreting Search Results	182
6.4.1	Viewing Match Info	183
6.4.2	Finding Similar Documents	183
6.4.3	Showing Document Parents	183

Chapter 1

Overview

The Thunderstone Search Appliance is a web walking and indexing device that allows a web site administrator to provide a high quality retrieval interface to collections of HTML and other documents. It is an application of Taxis and is written in Taxis's Web Script language named Vortex.

It consists primarily of the Taxis binary program and two Vortex scripts that are run by the Taxis CGI program on the appliance and are accessed from a web browser.

One script provides the administrative interface, another provides the site walker and indexer, and the third provides the search function that end users see.

1.1 Features

Here are some of its features:

- One or more web sites may be indexed into a single database.
- Multiple databases may be maintained.
- It supports cookies.
- There is support for meta data.
- It supports proxy servers.
- Robots.txt and meta robots are respected.
- It provides a totally customizable search interface.
- It provides a totally customizable site walker/indexer.
- A web site may be copied to the local file system.

There are many more features and options to tailor the Search Appliance's behavior to your needs.

1.2 Technical Support

Support for the Search Appliance is available via a searchable web message board. It is located at the following URL:

`http://thunderstone.master.com/texis/master/search/msgboard.html`

Anyone may read the discussions. To post a question or comment, you must create an account, which is free, and you must be logged in. Also, once you are signed up, you may “subscribe” to periodic email notifications of new postings to the board. You may select hourly, daily, or weekly notification of new postings.

If you subscribe to periodic notifications, and at some point in the future no longer wish to receive them, you may select “unsubscribe” again to enter the administrative area where you may delete your subscriptions.

Chapter 2

Installation

2.1 How to unpack and install the Search Appliance

For basic information about unpacking and installing the Search Appliance, refer to the Getting Started guide. This printed guide was shipped with the Appliance. In addition to the instructions it provides, it includes a sticker that lists important information unique to your Appliance. This information includes the original password and various network addresses.

2.2 Customizing the Search Appliance's Appearance

You may make common changes to the Search Appliance's search appearance by using `Search Settings` from the administrative interface main menu. You may select color, font, size, result style and order, as well as setting boilerplate HTML to wrap around the search form and results.

Chapter 3

Operation

3.1 Running the Administrative Interface

The Search Appliance's administrative interface is a web application that you access using your web browser. Access it using `http://YOURSERVER/telexis/dowalk`

Where YOURSERVER is the name (or IP address) of your Search Appliance.

When you run the administrative interface you will be asked for the login and password. By default there is one login name. It is `admin` in all lowercase. If no other accounts have been added, you will not have to enter the name. It will be filled in for you. Your login will be remembered in a cookie until you logout. This way, you don't need to enter the password every time you enter.

Note: If you share your computer with others, or it is available to people who should not be administering the Search Appliance, then you should logout when you are finished. This will help prevent unauthorized configuration of the Search Appliance.

The Search Appliance administrative interface uses JavaScript to enhance its functionality and make it easy to use, but the interface will also work well without JavaScript. No functionality of the Search Appliance will be lost if JavaScript is turned off in your browser (eg. to prevent pop-ups on other sites). In this document, the user interface description assumes that JavaScript is enabled.

3.2 First Time Run

A default password was setup for the Search Appliance which you should now enter at the prompt. If for some reason this step did not happen, the first time you run the administrative interface you will be asked to create and enter a password. You should choose a password that is easy for you to remember but hard for someone else to guess. You will need to enter the same password twice (two input boxes will be provided) to protect against typing mistakes. Passwords are case sensitive.

Once you create the password, you will be automatically logged in and shown the `Choose a profile` page. A default profile name and data directory will be filled in for you. You may change either of these if desired, then hit the `Create Walk` button. A new profile will be created but a site walk/index will not be

started yet.

You are then presented with the main walk settings page. The `Base URL` will be automatically filled in with the name (or IP address) of your web server. If you wish to walk a different site you may change the `Base URL` at this point.

If your site has pages that you want indexed, and these pages have extensions other than `.html`, `.htm`, or `.txt`, add the extensions to the `Extensions` list. Also note that extensions are case sensitive, unless you use `Ignore case` under `All Walk Settings`.

Once you're satisfied with the URL and extension settings, you may hit the `GO` or `Update and GO` button to begin a walk of your site. A walk will be started in the background and you will be taken to the `Walk Status` page. This page will show you the status of the walk in progress and indicate when the walk is complete. This page will automatically refresh every 10 seconds with the latest progress information until the walk is complete. When the walk is complete you will see a summary of errors.

Once the walk is complete, you may click `Live Search` on the menu at the top of the page. This will take you to the search that users will use. It is also the URL you can place on your web page(s) to send users to the search.

You now have a site index that you can use. There are many options to control the site walk as well as the search interface appearance. They are described in detail elsewhere in this manual. Use the `All Walk Settings` button on the administration script's menu to see all of the options. Click the question mark (?) next to an item to get help for that item.

3.3 Administrative Interface Overview

The Search Appliance's administrative menu has the structure given below. Each item is described on the pages that follow.

- Entry
 - Basic Walk Settings
 - Update
 - GO, Update and GO
 - STOP
 - All Walk Settings
 - Update
 - GO, Update and GO
 - STOP
 - Search Settings
 - Update
 - Best Bet Groups
 - List/Edit URLs
 - Walk Status
 - Refresh
 - STOP Walk
 - Query Log
 - Test Search
 - Live Search
 - Profiles
 - Create Walk
 - Select a Profile
 - Delete a Profile
 - Accounts
 - Add a User
 - Change Password
 - Delete
 - User Groups
 - Access Control
 - Maintenance
 - Documentation
 - Appliance Home
 - Logout

3.3.1 Entry

Upon entry to the Search Appliance's administration interface you are prompted for user name and password. If you have logged in previously and still have the cookie and have not logged out, the login page is bypassed and you are taken directly to Profiles (see section 3.3.11, p. 24).

Your login is remembered in a cookie until you logout. This way you don't need to enter the password every

time you enter. If you share your computer or it is otherwise available to people who should not be administering the Search Appliance, you should logout when you are finished.

3.3.2 Basic Walk Settings

This is the central area for configuring a walk. The most commonly used walk related options and their settings are presented and they may be changed here. The Basic Walk Settings are a subset of the All Walk Settings. Next to each option is a question mark (?) which, if clicked, takes you to help for that option. The options are documented individually later in this manual in section 3.4.

At the bottom of the page is a set of three buttons. Pressing any of the buttons affects all options on the entire page.

- **Update**
This button causes all changes on the form to be saved. No walk is started.
If the rewalk schedule has been changed, the new schedule will go into effect immediately.
If categories have been changed, the walk database will be updated to reflect the new categories. The search interface will reflect the new categories.
If single page, page file, or page URL has been changed, the listed individual pages will be fetched into the live search database and made available for searching.
If the word definition is changed, the search index on the live database will be dropped and recreated. Searches might not work while the index is being rebuilt.
- **GO or Update and GO**
The GO button will change to Update and GO after you make a change to any setting on the form. The ultimate behavior for either is the same.

The current settings from the form will be saved as is done when you click Update. Then a new walk will be started. The new walk will be performed to either a temporary database or the live database, depending on the setting of Rewalk Type (Section 3.4.11). Then you will be shown the walk status page where you may monitor the progress of the walk.

Changes to categories or word definitions will not be reflected until the walk finishes.
- **STOP**
When a walk is in progress the GO button is replaced by the STOP button. This button terminates the running walk and abandon the work that it has done so far.
- **Reset**
This button reverts all settings on the page to what they were when the page was first loaded.

3.3.3 All Walk Settings

This is the central area for configuring a walk. This is similar to Basic Walk Settings except that all walk related options and their settings are enumerated and may be changed here.

3.3.4 Search Settings

This page contains all of the settings related to the search interface that end users see when performing searches.

All search options and their settings are enumerated and may be changed here. Next to each option is a question mark (?) which, if clicked, opens help for that option. The options are documented individually later in this manual in section 3.6.

At the bottom of the page is a set of two buttons. Pressing any of the buttons affects all options on the entire page.

- **Update**

This button causes all changes on the form to be saved.

Changes made to the appearance options will be immediately visible in the test search. If apply appearance is checked, the changes will also be immediately visible in the live search.

- **Reset**

This button reverts all settings on the page to what they were when the page was first loaded.

3.3.5 Best Bet Groups

The Best Bets are grouped together. This allows different groups to be shown in different places, and easily rotated in or out. For example, you might have one group of links that you have determined to be the most probable results for a user's query, and another group that includes links you want to promote.

The Group Name is how the group will be identified elsewhere in the administrative interface. This should be chosen to readily remind you of the purpose behind the group.

The Result Type indicates which fields will be shown on the results page. The title and description are entered by the administrator, rather than always being taken from the page.

3.3.6 List/Edit URLs

On this page, you may list or delete all or selected URLs from the database. You should always list before you delete, so you know that you are deleting the correct ones. While listing URLs, you may display all known information about a given page. You may also create categories for selected sets of URLs from this interface.

If a walk is in progress, delete is disabled and you are given the choice of listing URLs from the live search database or the new database being built by the walk.

Select **List** or **Delete** from the drop down list. The default is always **List** for safety.

In the pattern box, enter the URL or pattern for URLs for which you want information. This may be an exact URL or a wildcard pattern, which lists all URLs matching the wildcard pattern. For a wildcard pattern, use asterisk (*) to match anything and question mark (?) to match any single character. You may enter up to 10 different URLs or patterns in the box to find them all at once. Put a space between patterns

when entering multiples. Leaving the pattern box blank implies *, and this will cause every URL in the database to be listed. Deletion will be denied if the pattern is blank or *.

Select the order in which you wish to see the list:

Depth	URLs encountered first in the walk will be listed first
URL	URLs are ordered alphabetically
Newest first	URLs are ordered by modification date with newest ones first
Oldest first	URLs are ordered by modification date with oldest ones first
Largest first	URLs are ordered by download size with largest ones first
Smallest first	URLs are ordered by download size with smallest ones first

Then `Submit`.

All matching URLs will be listed. Clicking on a listed URL opens a page of details about that URL. On that detail page, everything the database knows about that URL is presented. You can also see what pages refer to the selected page by clicking `Parents` and what pages the selected page refers to by clicking `Children`.

If your pattern matches less than the entire database, you will be given a form from which you can create a category using the same pattern(s). Simply enter the name of the category to create and click `Submit`. The name is the name that users will see on the search form. This new category will also appear on the main settings page along with the other categories. It will also be immediately available to search users.

Live Search Database and New Walking Database

These options are presented on the `List/Edit URLs` page (see 3.3.6) if a walk is active. They allow you to choose which database to query. The “Live” database is the one from a previous successful walk that is what search users see. The “New” database is the database currently being built by the new walk. It is not visible to search users.

3.3.7 Walk Status

This page shows the status of the latest walk for the current profile. If a walk is in progress, it is the one reported.

During an active walk, it indicates a summary of how many pages are to be walked in the next hour, how many were walked in the last hour, and the total number of pages. There is a list of the most-recent URLs fetched, with number of errors and duplicates found, followed by a list of the next URLs to be walked. Below that is summary information about the walk itself, including walk start time, starting URLs, and some profile settings. The Walk Status page updates automatically every 10 seconds until the walk is complete or another page is selected. (After 10 minutes of user inactivity it will refresh once a minute to save traffic.)

When no walk is in progress, the report also includes a list of errors and duplicates encountered. If the last walk was abandoned, the report includes information about how far it went, as well as the report from the last complete walk.

Now button

During the walk the **Refresh display: Now** button may be selected to force a Walk Status display refresh before the 10 second automatic refresh. Note that this only affects the display, not the walk itself.

Pause/Auto button

The **Refresh display: Pause** button pauses the Walk Status display (prevent the browser from refreshing the display every 10 seconds); this changes the button to **Auto** which will have the opposite effect (resume the auto-refresh). This is useful when examining the status page in detail, and avoiding being interrupted by the browser auto-refresh. Note that both buttons only affect the display, not the walk itself.

STOP walk button

The **Current run: STOP** walk button on the Walk Status page stops the current walk. If the walk type is **New**, the walk will be abandoned (current live search is left intact and not updated). If the walk type is **Refresh**, the new pages are always live (since refresh uses one database), but the search indexes are not updated.

Pause walk and Make live button

The **Current run: Pause walk and Make live** button pauses the current walk, updates its search indexes for speed, and makes the walk live (ie. deletes the current live database and replaces it with the current walk). This can be useful if you ran out of disk space while indexing and subsequently freed up some space, or if a long running walk was stopped and you want to use the incomplete walk. If the walk was abandoned due to an error, make sure you resolve the problem before trying to make the new database live.

3.3.8 Query Log

The query log pages provide detailed and summary information about queries. Query logging must be turned on to generate information on the query log pages. If query logging has never been turned on for the current profile, there will be nothing to see. The query log is erased each time the database is rewalked.

The pages are as follows:

- Query Report
- Top Query Words
- Top Queries
- No Hits
- Best Bet Clicks

The query log lists the time that each search occurred, the IP address of the web user performing the search, the number of hits for the search, and the user's query. For URL clickovers, it displays the query instead of the number of hits and the actual URL instead of the query.

Selecting the Date/Time for a listed query will display a page with complete information about the search. This page includes everything from the summary list, and any non-default parameter settings from the search. A hyperlink is provided so that you may perform the same query as the user.

3.3.9 Test Search

This hyperlink opens the search interface. It forces the interface to use the search settings listed on the `Search Settings` page, whether they have been applied or not. This allows you to test search settings without affecting end users until you are satisfied with the new settings.

This mode also places two extra hyperlinks at the top of the search pages. `Back to Administration` allows you to return to the Search Appliance administration interface. `Make this appearance live` does that too, but it additionally makes the search settings you are testing "live", so that end users also see the search setting effects.

3.3.10 Live Search

This hyperlink opens the Search Appliance search interface as end users see it.

3.3.11 Profiles

This page presents a list of existing profiles. A profile contains the walk and search settings for a collection of pages. You can click on the profile name to see and/or change its settings and status or to start a walk.

You can click on `Delete` next to a profile to delete that profile. You will be asked whether you really want to delete the profile or not.

When a profile is deleted, all of its settings are lost and any walk database it has created is deleted. There is no way to get back any of these items after the profile is deleted. You shouldn't delete a database that is being actively searched.

You may also create a new profile by entering a new name.

You can copy settings from an existing profile to your new profile by selecting its name from the drop down list. This allows you to set up another site similar to an existing one. It allows you to experiment with the walk settings for an existing site, without potentially harming the good walk that is being searched by your users.

3.3.12 Accounts

This section provides information to maintain multiple login accounts for access to the Search Appliance administration. All users are listed on this page. You may add users, delete users, and change individual

user passwords. The default user, called `admin`, may not be deleted.

The Accounts page also allows you to create multiple administrative users. There is no distinction among them after they are created. All users have full administrative permissions, and they may create and delete any user or change any user's password. This is a basic security mechanism meant to keep unauthorized persons from using the web based administrative interface. The purpose of supporting multiple administrative users is that you can create distinct passwords, which you can revoke in the future without needing to change a single global password that all administrators know.

The passwords are one-way (forward) encrypted. This means that a forgotten password may not be discovered. The only way to deal with a forgotten password is to change the password.

Add a User

To add an administrative user, enter the new user's login name and password. You will need to enter the new password a second time into the `Confirm` box to protect against typing mistakes (since you can't see the password you are typing).

Names and passwords are case sensitive. "Joe" is different than "joe". You should choose passwords that are easy to remember, but difficult for someone else to guess.

Change Password

Here you may change the password for the selected user. You will need to enter the new password twice to protect against typing mistakes (since you can't see the password you are typing). Enter the password once the `Password` box and again into the `Confirm` box

Passwords are case sensitive. "Joe" is different than "joe". You should choose passwords that are easy to remember, but difficult for someone else to guess.

Delete

This will delete the selected user. You will be prompted to confirm whether the user should really be deleted or not. Once a user is deleted, there is no way to get it back except to re-add it.

The default user, "admin", may not be deleted.

3.3.13 User Groups

User groups may be created on this page, by clicking the `Add a Group` link. Existing groups may be edited or deleted with the appropriate links. User groups are used to associate administrative users into similar-privilege groups for easier access control maintenance. See the User Groups section for more details (p. 65).

3.3.14 Access Control

The Access Control page allows configuration of administrative users' access to administrative actions (creating profiles, starting walks etc.). In conjunction with user groups, access control can be used to restrict certain users to only certain actions, instead of allowing all users access to all administrative functions. See the Access Control section for more details (p. 65).

3.3.15 Maintenance

The Maintenance page contains various links for maintaining and editing operating-system and overall settings.

See also Maintenance 3.11.

3.3.16 Documentation

This provides a hyperlink to the online version of this document.

3.3.17 Appliance Home

This provides a hyperlink to the online home of the Appliance.

<http://www.thunderstone.com/taxis/site/pages/Appliance.html>

3.3.18 Logout

This will log you out of the administrative interface and clear your login cookie. It then takes you back to the login page.

3.4 Basic Walk Settings

This page contains the settings that are used most commonly. They are available in Basic Walk Settings.

The settings on the Basic Walk Settings page are a subset of the settings on the All Settings page. Use the page that is most convenient for your current task.

3.4.1 Walk Summary

This is informational only. It contains summary information about the most recent walk and recategorizations. The information includes the date and time of the walk, whether the walk was successful, how many pages were indexed, and the number of duplicate pages.

3.4.2 Notes

This is a scratch pad area for the administrator of the profile. It in no way affects the walk or search.

3.4.3 Base URL

Syntax: one or more URLs, one per line

This is the address where the web crawler will start walking your site. If the whole site is to be searched, simply enter your web address, for example “`http://www.mysite.com`”. If the search is to be limited, specify the address to start the search or create a page listing the URLs to search. The search will only return information from your web site - no off-site searching will be done. Directory URLs should include a final forward slash “/”. Example - “`http://www.somehost.com/mysite/`”. If you have a virtual domain that just redirects to another URL, enter the destination URL as your Base URL instead of your virtual domain name.

You may specify multiple base URLs to index multiple sites; the Search Appliance’s idea of a “site” is a single host as identified by the hostname portion of a URL. Therefore `http://www.mysite.com`, `http://www2.mysite.com`, and `http://mysite.com` would all be considered different sites.

In version 4.02.1046373961 Feb 27 2003 and later, the special “protocol” `http-post` or `https-post` may be used for a Base URL. This uses the POST method instead of the GET method to fetch the URL, using the query string as POST data (it must be URL-encoded). This can be used to start walking at a login page form that requires POST instead of GET. Note that the URL stored in the `html` table will have the `-post` and query string removed for security. During a Refresh walk, when a URL is about to be refreshed, the probable Base URL that led to it (ie. the one with the longest prefix) will also be fetched. This helps ensure that login cookies are properly restored to allow the Search Appliance access during the refresh. Example:

“`http-post://www.somehost.com/login.asp?user=bigbird&pass=open-sesame`”

In version 5, a username and password may be given in the Base URL. Normally, if only one login is required to access the site to be walked, the username and password should be given in the Login Info walk setting. However, if several different logins are required, the additional logins can be specified as `user:password@` prefixed to the hostname in the Base URL. Note that the user/pass is for WWW Basic Authentication. If your site uses a custom or form-based login, use `http-post` instead. Example:

“`http://MyName:MyPassword@www.myhost.com/login.asp`”

See also URL file 3.5.6, URL URL 3.5.7, Single page 3.5.8, Page file 3.5.9, and Page URL 3.5.10 for more ways to specify URLs.

3.4.4 Enterprise

Syntax: a single domain name

The name of your company’s domain. This is useful if your company’s web presence consists of multiple hosts within its domain, and you want them all indexed together as a unit.

This allows you to walk any URLs encountered during the walk of the base site(s) that are within the given domain. The Search Appliance will attempt to guess this value for you, but you may set it to whatever you wish. Check the `Yes` box to enable this feature.

See also `Extra domains 3.5.13` which is the same but allows more than one domain. These options may be used together.

3.4.5 Robots

Syntax: select `Yes` or `No` buttons

robots.txt

With this set to `Yes`, the Search Appliance will initially get `/robots.txt` from any site being indexed and respect its settings for what prefixes to ignore. Ignoring `robots.txt` is not generally recommended.

See also `Robots.txt 4.5`.

Meta

Respect the meta tag called `robots`. With this set to `Yes` the Search Appliance will process and respect the robot control information within each retrieved HTML page.

See also `Robots.txt 4.5`.

3.4.6 Extensions

Syntax: one or more file extensions separated by space

A list of the URL extensions that the crawler will accept. The defaults are

```
.html  
.htm  
.txt  
.pdf
```

To search MS-Word documents, use `.doc`. For Shockwave/Flash use `.swf`. For WordPerfect documents specify whatever extension you use and ensure that the web server returns the MIME type `application/wordperfect` as there is no consistent extension for WordPerfect documents. Any extensions not listed here will not be searched or walked.

A few other extensions you may find useful are

```
.asp  
.cfm  
.jsp  
.shtml  
.jhtml  
.phtml
```

3.4.7 Exclusions

Syntax: zero or more strings, each on a separate line

Excludes URLs containing any of the specified literal strings anywhere in the URL (hostname, path, or query).

See also `Exclusion REX` 3.5.16 and `Exclusion prefix` 3.5.17 for more ways to exclude URLs.

3.4.8 Crawl Delay

Syntax: a whole number from 0 to 10

Causes the Search Appliance to wait the specified number of seconds between page fetches. Normally set this to 0, but increase it if the web server cannot handle being hit rapidly. Increasing this value forces the walk to take at least the following number of seconds to complete: the Crawl Delay number times the number of pages on the site.

Note: Using a delay larger than 0 forces `Threads`(3.4.9) to 1. A delay defeats the advantage of multiple threads and large delays could cause unexpected page fetch timeouts.

3.4.9 Parallelism

Syntax: whole numbers from 1 up

Threads

This is the maximum number of simultaneous page fetching threads to allow against each site. Setting `Threads` higher than 5 is probably not very helpful, unless you have many “Single Pages” that are on various hosts.

Servers

This is the maximum number of different web servers to walk simultaneously. Setting this too high can stress your memory, cpu, and network.

3.4.10 Verbosity

Syntax: whole number from 0 through 4

Sets how much information the walker should provide about what it’s doing. The default verbosity level is 2. The values are described in the following table.

The levels are cumulative. In other words, each level includes the previous levels.

3.4.11 Rewalk Type

Syntax: select from drop down box

Table 3.1: Verbosity Levels

Level	Description
0	Issue no messages except errors
1	Display starting point URLs
2	Display selected setting info
3	List URLs found in URL files
4	Indicate why URLs are rejected

This determines how rewalks are performed.

New

The type `New` creates a new database and does a complete walk of everything, starting with the Base URLs. A `New` walk does not disturb the existing database.

Refresh

The default rewalk type `Refresh` updates the existing database, and only downloads files that have been modified or created since the last walk. Pages that are no longer present on the server are removed from the database.

Here are other considerations for using `Refresh`. Pages that were referenced but were missing in the initial walk (the walk prior to the `Refresh`), but were added after the initial walk, will be missed by `Refresh` if their parent page has not been modified. If you change your settings to be more inclusive (ie add extensions, ignore robots, add domains, etc.), you should do a `New` walk once, because a `Refresh` is not likely to find the newly allowed data, unless all of the pages leading to this data have been modified.

If more than 30%-50% of your site changes between walks you may be better off using a `New` walk instead of `Refresh`. Also, many dynamic content generators do not give modified dates which will cause every page to be rewalked. In that case you should use `New` instead of `Refresh`.

Refresh in version 5 vs. 4

In the Search Appliance version 4 and earlier, the refresh walk checked every page in the database to determine whether it needed updating. Since only changed pages need updating, and those are typically a small percentage of the site, checking for changed pages is faster than doing a complete new walk. However, it is still time-consuming, because the web server must be accessed for every page on the site, and only the web server can inform the Search Appliance whether the page has changed.

In the Search Appliance version 5 and later, there is an improved refresh process. The walk is adapted to focus on the small but important group of changing pages. As each page is walked, the Search Appliance calculates a refresh period for that individual page. The calculation is based on whether the page has changed since the last time it was fetched, and how long ago that fetch was. This refresh information is used

to determine when the page should be checked again. In this way, the walk prioritizes the walking of pages that change often or are new, and it delays the fetch of pages that seldom change.

Thus, when a walk (scheduled or manual) takes place, only the pages that need to be refreshed now are actually fetched – not the entire database. The result is a database that is updated by a process that consumes fewer server resources.

Rewalk Type Summary Table

The following table summarizes the trade-offs for the new and refresh rewalk types.

Method	Advantages	Disadvantages
New	Guarantees most accurate representation of current site. Does not disturb live search database.	Uses more bandwidth and temporary disk space. Longer time before site changes are reflected in live search.
Refresh	Faster. Uses less bandwidth and temporary disk space. Site changes are reflected in live search much sooner.	Could get out of sync with actual site under rare circumstances. A lot of changed pages could substantially slow searches during the walk. Requires If-Modified-Since support on walked web server.

3.4.12 Rewalk Schedule

Syntax: select from drop down boxes

This performs a rewalk on the schedule specified. The rewalk action is the same as the one that can be started manually by clicking the GO button. The `Frequency` defines how often to automatically rewalk. The `Hour` defines which hour to start the rewalk for daily or weekly runs.

See also `Notify` 3.5.2. If you are using “On Change” see also `Watch URL` 3.5.1.

3.4.13 Action Buttons

These buttons tell the Search Appliance to do something now. Only the buttons applicable to the current status are displayed. The buttons are as follows:

- `Update`: Save the current settings for future use but don’t begin a walk.
- `GO`: Begin a walk using the current settings.
- `Update` and `GO`: Save the current settings then begin a walk using those settings.
- `STOP`: Stop and abandon the walk that is currently running.

See the Walk Settings section (3.3.2) for details about the operation of these buttons.

3.5 Advanced Walk Settings

These are the advanced settings that are used less commonly than the settings available in Basic Settings. The advanced settings are available in All Walk Settings.

3.5.1 Watch URL

Syntax: an HTTP URL

The URL specified here will be refreshed every time that The Search Appliance starts a refresh walk. This can be used if you have a page that lists new documents that are added to the site as it will ensure that the links are found as soon as possible.

3.5.2 Notify

Syntax: an email address

If this is set, a summary report will be sent to the supplied email address when a scheduled rewalk occurs.

3.5.3 Attach Logs

This selects the log files to attach to the walk notification. The log files and walk errors are for the period of the refresh walk, and are sent as tab separated files that can be opened with programs such as Excel for further processing.

3.5.4 Categories

Syntax: textual name and URL pattern pairs, additional input boxes will appear as you fill the ones provided

The Search Appliance can create searchable sub-categories that will appear in a drop down box on the Search page. Enter the name of the category on the left, and its corresponding URL pattern on the right. URL patterns may contain asterisk(*) to indicate “anything” and question mark(?) to indicate any single character. There may be more than one pattern for each category. Separate multiple patterns with space. The following table provides an example.

Table 3.2: Example Categories

Category	URL Pattern
Demonstrations	<code>http://SERVER/demos/*</code>
Manuals	<code>http://SERVER/manual/*</code>
Books	<code>http://SERVER/a1/* http://SERVER/b3/*</code>

This example would create a category named “Demonstrations” which would only search the URL “`http://www.mysite.com/demos/`” and any files under this directory, thereby creating a more concise match to the users search. The same is true for “Manuals”. The “Books” category would include

pages from both the “a1” and “b3” directories. The user would now have the option to search within just these categories or the entire database. The pattern should NOT be a single page unless you want a category with a single page in it (e.g. `http://www.mysite.com/manual/index.html` would be incorrect). It should typically be a prefix for a directory that has multiple pages within it followed by an asterisk (*).

3.5.5 DBWalker

Here you can select one more more database walking configurations to include in this profile. This can be done in addition to specifying any Base URLs (section 3.4.3). To select multiple configurations, hold `Ctrl` while clicking in the select box.

For more information on the database walker module, please see the DBWalker section (4.19, pg. 103) of the manual.

3.5.6 URL File

Syntax: the full path to a file on the web server’s disk

This allows you to specify a file containing a list of site URLs to walk. This is an additional way of specifying more Base URLs 3.4.3. This file will be reread each time a rewalk is started. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

3.5.7 URL URL

Syntax: an HTTP URL to a plain text file (NOT HTML)

This allows you to specify the URL of a plain text file containing a list of site URLs to walk. This is an additional way of specifying more Base URLs 3.4.3. This URL will be refetched each time a rewalk is started. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

3.5.8 Single Page

Syntax: one or more HTTP URLs, one per line

Here you may specify URLs for individual pages to include in the index. These pages are fetched and stored in the database like others but the hyperlinks on them are not followed during a walk.

If you change this and click “Update” instead of “GO” the added pages will be fetched immediately and added to the existing database. Pages removed from the list will NOT be removed from the database until the next rewalk.

3.5.9 Page File

Syntax: the full path to a file on the web server’s disk

This may be used to specify a file containing URLs for individual pages.

If you change this and click “Update” instead of “GO” the added pages will be fetched immediately and added to the existing database. The file itself is not checked for changes, and pages removed from the file will NOT be removed from the database until the next rewalk. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

See also `Single` page 3.5.8.

3.5.10 Page URL

Syntax: an HTTP URL to a plain text file (NOT HTML)

This may be used to specify the URL for a plain text file containing URLs for individual pages. In the file, the list of URLs can be one URL per line (preferred) or delimited by any number of spaces.

If you change this and click “Update” instead of “GO” the added pages will be fetched immediately and added to the existing database. The file itself is not checked for changes, and pages removed from the file will NOT be removed from the database until the next rewalk.

See also `Single` page 3.5.8.

3.5.11 Strip Queries

Syntax: select Yes or No button

Strip query strings from all URLs. Some URLs have query strings on the end indicated by a question mark (?). With this option set to Yes, all query strings are removed from URLs before they are processed or retrieved.

3.5.12 Ignore Case

Syntax: select Yes or No button

This tells the Search Appliance whether to ignore case in URLs or not. The case of hostnames is always ignored but the case of paths and filenames is respected. Some web servers don't respect case and people use various random capitalizations within filenames making the same file look like different URLs.

3.5.13 Extra Domains

Syntax: one or more domain names separated by space or line break

Allow walk to fetch pages from any host in the specified domain(s). Any URL with a hostname ending in any of the specified domains will be accepted.

e.g.: Given a base URL of `http://www.mysite.com/` and extra domain `othersite.com` the Search Appliance will walk all of `www.mysite.com` and any URLs referring to any machine in

othersite.com.

This option is not a “restrictor” but an “enabler”. All hosts specified will be walked and any others that match the given domain(s) will also be walked.

Note: This option does NOT direct the walk to completely index every web server in the specified domain. It simply allows walking them if a reference to them is encountered.

3.5.14 Extra Networks

Syntax: one or more IP address prefixes separated by space or line break

Allow walk to fetch pages from any host within the network specified by the numeric IP address(es).

e.g.: Given a base URL of `http://www.mysite.com/` and extra network `192.0.2` the Search Appliance will walk all of `www.mysite.com` and any URLs referring to any machine having an IP address prefix matching `192.0.2`.

Note: This option does NOT direct the walk to completely index every web server in the specified network. It simply allows walking them if a reference to them is encountered.

Note: Using this option has the potential to slow the walk, because every URL’s hostname must be looked up. If there are many different off-site hosts, or your DNS is slow, the walk may be slowed substantially.

3.5.15 Extra URLs REX

Syntax: zero or more regular expressions (REX), separated by space or line break

Allow walk to fetch additional URLs matching any of the specified regular expressions anywhere in the URL (hostname, path, or query). Most commonly used in conjunction with a hostname to fetch matching URLs on an additional host. Links still need to be found to those pages for them to be indexed.

Available from version 4.3.9.

Table 3.3: Extra URLs REX examples

REX	Matches
<code>>>http://library/!=!math*math=</code>	Urls on the server library containing “math”

See also `Extra Domains` 3.5.13.

3.5.16 Exclusion REX

Syntax: zero or more regular expressions (REX), each on a separate line

Excludes URLs matching any of the specified regular expressions anywhere in the URL (hostname, path, or query).

See also `Exclusions` 3.4.7, `Exclusion prefix` 3.5.17 and `Exclude by Field` 3.5.18.

Table 3.4: Exclusion REX examples

REX	Matches
/scratch[0-9]/	a subdirectory named “scratch” followed by a single digit
[^\alnum]test[^\alnum]	the word “test” (but not retest or tester etc.)

3.5.17 Exclusion Prefix

Syntax: zero or more URL prefixes, each on a separate line

Excludes URLs beginning with any of the specified prefixes. The entire URL (hostname, path, and query) is used for comparison.

Examples:

```
http://www.mysite.com/scratch0/
http://www.mysite.com/scratch1/
http://www.mysite.com/books/t
```

See also Exclusions 3.4.7, Exclusion REX 3.5.16 and Exclude by Field 3.5.18.

3.5.18 Exclude by Field

Syntax: Metamorph query, field to search, what to exclude

This provides more flexible control of what to exclude and how to exclude it. One exclusion per row of controls may be entered; new blank rows will be provided as rows are used. The Query column is where a Metamorph query (ie. a typical search on the Search Appliance) is entered: eg. several keywords or a regular expression. The Meta and Field columns determine what the Query searches: if Meta is non-blank, that named meta field is searched, otherwise the field selected in Field is searched. The Exclude column controls what happens for pages that match: Pages and links indicates that both the matching page and its links are to be excluded; Pages only indicates that the matching page is to be excluded but its links are still followed – this is useful for excluding navigation-only pages; Links only indicates that the page is still included but its links are excluded.

See also Exclusions 3.4.7 and Exclusion REX 3.5.16.

3.5.19 Additional Fields

Syntax: Name, Type, Searchable, Sortable, Output

The additional fields allow you to add up to three additional fields to the index, which can be included in the output if you use the XSL or XML output, sorted on, and searched on.

The Name specifies the name of the element that will hold the field contents in XML.

The field can be populated with the Data from Field (section 3.5.20) settings.

3.5.20 Data from Field

Syntax: REX expression, field to search, what to exclude

This provides alternate means of setting the `Modify Date`, `Title` or `Description` fields for searching. It allows getting page information from non-default places by searching and optionally replacing the data. One inclusion per row of controls may be entered; new blank rows will be provided as rows are used. The `Search` column is where a REX expression must be entered. If you want to match the entire field use `.*` as the expression if you want to override the default only if a value exists, or `.*` to override with an empty value if no value is specified. The `Replace` column is used (optionally) to replace the data matched by the search. The `Meta` and `Field` columns determine what the `Query` searches: if `Meta` is non-blank, that named meta field is searched, otherwise the search field selected in `Field` is searched. Which `Field` controls the destination for the data (what it sets).

3.5.21 Required REX

Syntax: zero or more REX expressions, separated by whitespace

If specified, *all* URLs walked by the Search Appliance must match at least one of these expressions. Opposite of `Exclusion REX`.

3.5.22 Required Prefix

Syntax: zero or more URL prefixes, separated by whitespace

If specified, *all* URLs walked by the Search Appliance must match at least one of these prefixes.

3.5.23 Max Page Size

Syntax: a whole number from 1 up

Sets retrieved page size limit to the specified number of bytes. Pages larger than the limit will be truncated - not discarded.

Note: PDF files tend to be very large for the amount of text contained within them. Truncated PDF files are not processable due to their design. Make sure this setting is large enough to handle the largest PDF file you want to index.

3.5.24 Max Pages

Syntax: a whole number from -1 up

Limits the number of pages retrieved in a run to the specified number. Use -1 for no limit.

3.5.25 Max Bytes

Syntax: a whole number from -1 up

Limits the number of bytes retrieved in a walk to the specified number. Use -1 for no limit. The actual limit is rounded up to include the size of the last page so that it does not get truncated.

3.5.26 Max Depth

Syntax: a whole number from -1 up

Limits the depth of page retrieval to the specified number. Use -1 for no limit. Depth is determined by counting how many links were traversed to reach a particular page. The base URLs are all at depth 0. URLs referred to by the base URL are depth 1, and so on.

3.5.27 Page Timeout

Syntax: a whole number from 1 up

Causes the Search Appliance to timeout after the specified number of seconds during each page fetch. This includes the time to lookup the IP address of the host, make the connection to the server, and download a single page. A timeout does not cause the entire process to quit. That page is just skipped and considered unavailable.

3.5.28 Meta Tags

Syntax: zero or more meta tag names, each on a separate line

This option tells the Search Appliance to look for the specified meta data in fetched documents and store it in the database. Then, this data is included in text searches. The meta tags “Description” and “Keywords” do not need to be specified here because they will be indexed by default. See below.

3.5.29 Standard Meta

Syntax: select Yes or No button

This option indicates whether to automatically extract the standard meta tags “Description” and “Keywords” from HTML documents. If “Yes”, description and keywords meta data will be extracted and stored in their own fields within the database, unlike other meta data which will be collected and placed together into a single meta field in the database. These meta tags will be included in the search with a higher precedence than other meta tags.

3.5.30 All Meta

Syntax: select Yes or No button

Extract all meta data from HTML documents and place this data into the meta field for searching. This eliminates the need to know the name of all possible meta tags, but it also opens the possibility of recording all manner of nonsensical meta data.

3.5.31 Storage Charset

Syntax: standard IANA character set (charset) name

This sets the charset for storing page text in the database during walks. Pages will be translated to this charset when inserted. If a page cannot be translated, it is stored and labelled with its source charset (if known). If left empty (the default) it is UTF-8. This charset should be a superset of US-ASCII (same 7-bit sequences), and translatable by the Search Appliance from all walked pages' source charsets.

Note that this is *not* necessarily the charset that search results will be displayed in: see Display Charset under Search Settings. This setting is the default value for Display Charset; see notes under Display Charset.

3.5.32 Source Default Charset

Syntax: a standard IANA character set (charset) name

If the source charset for a walked URL is not labelled and cannot be determined, assume it is this character set. Default is ISO-8859-1. This should only be changed if a large number of walk pages are in an unlabelled different charset, eg. a Windows charset.

3.5.33 XML UTF-8

Syntax: select Yes or No button

Whether to attempt to clean up UTF-8 data for XML output: remove invalid sequences and characters. Should be Yes if XML output (eg. result style 8) is used (and Storage Charset should be empty). This helps avoid browser errors with XML pages. *Note:* if XML output is *not* being used, this should be set to No, as certain characters that are HTML-safe but not XML-safe will be removed if enabled.

3.5.34 Keep HTML

Syntax: select Yes or No buttons

Specifies whether to include the named type of text in the database.

ALT text

ALT text from IMG or AREA tags.

<STRIKE>

Text between <STRIKE> and </STRIKE> tags.

Text between and tags.

<FORM>

Text between <FORM> and </FORM> tags.

3.5.35 Keep Links

Syntax: select Yes or No buttons

Specifies whether to follow the named type of links when crawling.

Stylesheet

Links from <LINK HREF=... REL=stylesheet> tags. Note that non-stylesheet <LINK> tags will still be followed. The default is N.

<FORM>

Links from <FORM ACTION=...> tags. Without the rest of the form properly filled out, such links can often produce nuisance error pages from database-driven sites. The default is N.

3.5.36 Remove Common

Syntax: select Yes or No button

This causes common leading and trailing text from pages to be removed from the database. This is good for eliminating navigation menus and other static boilerplate text at the beginning and/or end of each page.

3.5.37 Ignore Tags

Syntax: one or more pairs of strings, more input boxes are added as you fill string pairs

All data between the specified begin and end will be stripped from the HTML before the text is extracted. These are simple strings, not patterns or REX's and the case is ignored. This is useful for excluding boilerplate or otherwise unwanted portions of HTML documents.

3.5.38 Keep Tags

Syntax: one or more pairs of strings, more input boxes will be added as you fill string pairs

All data NOT between the specified begin and end will be stripped from the HTML before the text is extracted. These are simple strings, not patterns or REX's, and the case is ignored. This is useful for extracting prime interest areas of HTML pages without the surrounding boilerplate.

3.5.39 Ignore Characters

Syntax: List of characters

List characters here which should be removed from the text and query. These can be punctuation that is optional. Examples are part numbers, phone numbers, etc. Take care to avoid removing important characters, which you may want to delimit words. Eg. with the setting “-@”, the text “part 123-45@6” would be stored (and searchable as) “part 123456” instead.

3.5.40 Plugin Split

A group of settings that control whether and how to split `anytotx` plugin output into multiple sub-URLs in the table. Non-text files, such as PDFs, that `anytotx` processes are often very large or composed of sub-files. The Plugin Split setting allows these files to be split up for finer-grain searching. Split files will cause more than one URL to be entered in the `html` table (and thus also in potential search results) for the original URL. Such subsequent URLs will have an anchor appended to distinguish them from each other; usually this is the sub-file name, but it may be generic eg. “#part5” if there are no sub-files. *Note:* adjusting any of these settings can affect the ability of Refresh-type rewalks to complete successfully (New walks operate as usual).

Depth The Depth setting controls at what depth to split `anytotx` output. Each time a multi-file archive is unpacked by `anytotx`, the depth increases. Depth 0 (the default) means split at the top level (ie. do not split). Depth 1 would therefore insert each file of a ZIP file as a separate URL in the table.

Bytes The Bytes setting controls how many bytes each part will be after the file has been split. The default of 0 indicates do not split. This is useful for large monolithic files that have no detectable sub-file or page structure. If both Pages and Bytes are set, the first limit reached is used for each part.

AtPage The AtPage setting controls whether to force the Bytes-controlled splitting to occur at a page boundary (a Ctrl-L). Checking this may make each part arbitrarily larger than the Bytes setting, because a part may extend to the next page break. With this setting unchecked, a part may be up to 50% larger than the Bytes setting, because the page-break check will only go that far over the limit.

Pages The Pages setting controls how many pages to group in a part. The default of 0 does not split at all. If both Pages and Bytes are set, the first limit reached is used for each part. For example, setting Pages to 10 and Bytes to 100000 would break at 10 pages or 100KB, whichever comes first. This is useful to catch page-bounded documents like PDFs, and simultaneously avoid generating huge text for non-paged documents.

Plugin Split was added in version 4.03.1049838346 Apr 8 2003.

3.5.41 Word Definition

Syntax: one or more regular expressions (REX), each on a separate line

Sets the word matching expression(s). Each line is a regular expression defining what is considered a word within the textual content of the retrieved documents during the index process. The default expressions index normal words and some special items such as domain names.

You may supply multiple expressions, one per line, if you can't define your idea of all possible words in one expression.

For example, `>>\alpha=\alnum{1,20}` will index “words” beginning with an alphabetic character followed by 1 to 20 alphabetic or numeric characters.

Changing the word definition with `Update` instead of `Update` and `GO` will cause the existing search index on the data to be dropped and rebuilt. The database will not be searchable during the short time that the index is being rebuilt.

3.5.42 Index Fields

Syntax: list of fields ordered by desired weight

These fields will be searched by the user's query. Fields listed higher will be weighted higher in search results, according to the Position in Text search setting.

Note that changing these fields will cause indexes to be rebuilt, which may take several minutes or more for large profiles, and the old setting will be used until the index rebuild is complete.

3.5.43 Primer URL

Syntax: type, optional URL, optional checkbox

The `Primer URL` will be fetched before starting a crawl. It will not be stored in the search database, but is instead used to “prime” the Search Appliance with any necessary credentials (eg. cookies) for accessing the rest of the site. By default, the Base URL is used, in case any session/ASP cookies are needed.

If a form-based login must be filled out before accessing a site, the `Primer URL` can be set to the `<FORM ACTION> URL` of the login (fully-qualified), with any form variables (eg. user/pass) filled out in the query string. If the `<FORM METHOD>` must be POST instead of GET, the URL protocol may be changed to the pseudo-protocol “http-post”. Eg.:

```
http-post://login.acme.com/login.asp?User=Admin&Pass=open-sesame
```

would be submitted using the POST method, with the given query-string variables sent as the content. Note that the query-string variables and values should be URL-encoded.

For HTTP Basic or NTLM protected web sites, the `Login Info` setting should be used instead.

3.5.44 Login Info

Syntax: name and password

Specify a username and password for sites that require a login to view certain pages. These are used with HTTP Basic, Windows NTLM, and FTP authentication. Other authentication methods are not supported currently. Without proper login, protected pages will be skipped.

If you are trying to walk a site where a login form is provided on a web page, you may be able to walk it by using the action URL from the form with the form variables encoded onto the end as your base URL. For example if the form variable names were Uname and Upass and the action URL was `http://www.mysite.com/login.asp` you may be able to use a URL like `http://www.mysite.com/login.asp?Uname=YOURNAME&Upass=YOURPASSWORD`

Note: The search interface displays hit context and has an option to view the entire text of the page. This allows search users to view “protected” pages without entering a password.

3.5.45 Proxy

Syntax: the full URL to a web proxy server

This specifies the URL (not just hostname) of a proxy web server through which to pass page fetch requests. Blank means don't use a proxy.

3.5.46 Proxy Login Info

Sets the user name and password to authenticate to proxy servers, using the Proxy-Authenticate header and Basic Authentication. Used if the Proxy URL is filled in. Added in version 4.01.1031600000 Sep 9 2002.

3.5.47 Cookie Source Path

File path to a Netscape or Microsoft Internet Explorer format cookie file to read at start up. This allows persistent cookies saved by a browser to be read by the Search Appliance, so it can inherit the browser's state. To easily walk a site that requires a custom login (ie. not HTTP Basic authentication), and that uses persistent cookies, just login normally using a browser run *on* the Search Appliance machine itself. Then, enter that browser's cookie file in the Cookie Source Path setting (this is typically `%USERPROFILE%\Cookies` for Explorer on Windows). Then, The Search Appliance will automatically inherit the browser's permissions. Added in version 4.02.1042043803 Jan 8 2003.

3.5.48 Off-Site Pages

Syntax: select Yes or No button

Allow retrieval of individual off-site pages. By default the Search Appliance will not retrieve pages that are not on the same host as the base URL(s). Using this option, pages not on the same machine will be retrieved, but none of the pages that they reference will be walked. This option also allows off-site redirects, frames, and iframes to be fetched.

3.5.49 Stay Under

Syntax: select Yes or No button

When this flag is Yes, walks will stay under the directory specified in the base URL(s). When this is No, if a hyperlink to another location on the same site is encountered, the will follow the link. In neither case will the walk go to other sites unless they are in the list of walk URLs or allowed domains or networks.

3.5.50 Prevent Duplicates

Syntax: select Yes or No button

This option enables extra checking for duplicate documents. Documents with the same content are only be stored once, even if their URLs are different. This is accomplished by hashing the textual content of the page and not storing any page with a hash code that is already in the database.

3.5.51 Duplicate Check Fields

Syntax: checkboxes to choose fields

These are the fields which will be checked for duplicate prevention (if `Prevent Duplicates` is enabled). The concatenation of these fields is hashed for each incoming document, and if the hash is the same as an existing document, the incoming document will be discarded as a duplicate.

By default, only `Body` is checked, as the body is the majority of search content for a document, and thus another document that has an identical body should be considered a duplicate even if it has a slightly different title or description.

However, sometimes errors in processing (eg. `anytotx`) can cause the bodies of large numbers of documents to become empty and thus be considered duplicates of each other and removed. In this case it may be desirable to either turn off `Prevent Duplicates` or check more fields in `Duplicate Check Fields`.

Note: Changing `Duplicate Check Fields` after a walk has completed (ie. before a later `Refresh` type walk) may cause new documents to not be removed as duplicates as expected, since the pre-existing documents' hashes are now for a different set of fields. This will not cause errors or corruption; it just might leave some newly-duplicate documents in the database.

3.5.52 All Extensions

Syntax: select Yes or No button

Retrieve all files instead of only those listed in `Extensions`. This turns off checking of URL extensions. All URLs will be retrieved regardless of the extension (including images and such files).

3.5.53 Store Refs

Syntax: select Yes or No button

Controls whether URLs referenced by retrieved pages are added to the refs table. This can save some time

during the walk, as well as, disk space if it's turned off. But turning it off prevents the "Show Parents" option in the search from working. It also reduces the detail available from walk error reports.

3.5.54 Inline Iframes

Syntax: select Yes or No button

This indicates whether to treat iframes as a part of the page they are on or as separate stand alone pages. Selecting Yes will make them part of the page. Selecting no will make them separate.

3.5.55 Max Frames

Syntax: a whole number from 0 up

This indicates the maximum number of frames allowed on a page. Pages with more frames than this are discarded. If this is set to 0, the frames of framed documents are treated as independent, stand-alone pages.

3.5.56 Execute JavaScript

Syntax: select Yes or No button

Execute JavaScript that is contained on fetched pages and that might alter or generate the page content and URLs.

3.5.57 Fetch JavaScript

Syntax: select Yes or No button

Fetch JavaScript that resides at a separate URL instead of being inline on the page (eg. `<SCRIPT SRC>` tags).

3.5.58 JavaScript String Links

Syntax: select appropriate checkboxes

Sets which additional sources of potential JavaScript links to check. Some JavaScript links may not be found when scripts on a walked page are executed, so the internal list of all JavaScript string objects is scanned for potential URLs according to the checked boxes. `Menu` will look for common JavaScript menu navigation system links; `Protocol` will look for strings that look like valid fully-qualified Web links; `File` will look for probable file strings.

Note that any of these sources may potentially find incorrect links, especially the `File` type. Checking `File` is generally used only as a last-ditch effort to find some JavaScript links.

3.5.59 Debug JavaScript

Syntax: select Yes or No button

Print additional debugging messages for JavaScript errors.

3.5.60 Protocols

Select which protocols to allow to be fetched. If a protocol is not enabled, but the Base URL uses it, it will be automatically enabled for the walk. The protocols currently supported are `http`, `https`, `ftp`, `gopher` and `file`.

3.5.61 Embedded Security

Select the security for embedded objects on a page (eg. frames, scripts). Any fetches any required object. `Non-decreasing` will fetch a required object if its security (`https://` vs. `non-https://` in the URL) is not less than the main page, ie. an `https://` object on an `http://` page will be fetched, but not vice-versa. `Non-increasing` is the opposite. `Same protocol` requires that the protocol of the object be the same as the main page.

3.5.62 Entropy Source

Selects standard (default) or alternate entropy source. Entropy is used to initialize the SSL/https plugin. The standard sources should be sufficient; the alternate source is only needed if the `prngd` daemon (some Unix platforms) is required but cannot be successfully run. *Note:* Setting the source to Alternate will decrease SSL/https security.

3.5.63 Max Redirects

Syntax: a whole number from 0 up or -1

This indicates the maximum number of redirects that are followed when attempting to retrieve a page. If set to -1 then redirects will not be followed when attempting to retrieve the page, but will be treated as a link.

3.5.64 Index Name

Syntax: one or more filenames separated by space

Set the filename assumed for directory URLs. The default is `index.html` and `index.htm`. This filename will be removed from stored URLs to prevent redundant fetches of the page. So the URLs `http://www.mysite.com/fun/` and `http://www.mysite.com/fun/index.html` will be considered the same and only be fetched once (as `http://www.mysite.com/fun/`).

3.5.65 DNS Mode

Syntax: choose from drop down list

This controls how the Search Appliance looks up IP addresses for hostnames. “Internal” uses Taxis’s own internal parallelizing name lookup routines. “System” uses the standard system routines. You should use “Internal” unless it causes compatibility problems.

3.5.66 User Agent

Syntax: full user-agent string

Set the User-Agent (browser type) to report to web servers. Normally the Search Appliance reports itself as Mozilla version 4.0. Modify this setting to report as a different user agent. If you want to emulate a particular browser, you can access your site with that browser, then check the site’s transfer log to see what user agent string was logged (typically the last double-quoted entry on the line).

3.5.67 Mime Types

Syntax: one or more acceptable MIME types, each on a separate line

These are the Multipurpose Internet Mail Extensions (MIME) types that the Search Appliance informs the web server are acceptable. MIME types have the syntax `type/subtype`. Either `type` or `subtype` may be `*` to mean “any”. By default all MIME types are allowed (`/*/*`).

3.5.68 Respect Expires Header

Syntax: choose from drop down list

For `refresh`-type walks, this controls how the Expires header is used. Set to `No` the Expires header will be ignored. Set to `Limited` the Expires header will be used, but limited by the Minimum and Maximum Refresh Times. Set to `Yes` the Expires header will be treated as definitive.

Invalid and out of range headers will be ignored, with the exception of “0”.

3.5.69 Default Refresh Time

Syntax: choose from drop down list

For `refresh`-type walks, this is the default time period to initially try refreshing a URL; typically set to 1 minute. Note that the actual refresh period is dynamically computed for each URL based on how often it changes.

3.5.70 Minimum Refresh Time

Syntax: choose from drop down list

For `refresh`-type walks, this is the minimum time period to try refreshing a URL. The actual refresh period is dynamically computed for each URL based on how often it changes, and will not be less than this value. This prevents too much time being spent refreshing a very dynamic page (ie. constantly refreshing it and loading the web server). Typically set to 1 minute.

3.5.71 Maximum Refresh Time

Syntax: choose from drop down list

For `refresh`-type walks, this is the maximum time period to try refreshing a URL. The actual refresh period is dynamically computed for each URL based on how often it changes, and will not be greater than this value. This ensures that all URLs – even relatively static ones – are eventually checked for changes.

3.5.72 Maximum Process Size

Syntax: choose from drop down list

Upper limit to memory size of walker processes. If a walker process exceeds this limit, it is re-started (at the same point it left off) by the dispatcher, at most once. If the same child repeatedly exceeds this limit, the walk may stop until it is re-started via schedule or manually.

3.5.73 Replication Settings

Syntax: List of hosts and profiles

A list of hosts and profiles to send walk data updates to. The hosts must have the sending server listed as a cluster member under the system-wide settings.

See also “Replication” 4.17.

3.6 Search Settings

This group of options applies to the standard search and provides a convenient way to make common changes to the search behavior and appearance.

See also “Customizing the Search Appliance’s Appearance” 2.2.

3.6.1 Notes

This is a scratch pad area for the administrator of the profile. It in no way affects the walk or search.

3.6.2 Query Logging

Syntax: select Yes or No button

This indicates whether the search should log user queries. If Yes, users' queries are logged to the querylog table of the database. The contents of this table may be viewed from the Query Log menu of the Administrative Interface.

Note: The query log table gets erased during every new walk. You will only be able to view queries that have occurred since the latest new walk. Refresh walks do not cause the table to be erased.

3.6.3 Rotate Schedule

Syntax: The day of week (or daily) and the time of day to rotate

This selects when to rotate query logs on this profile. During a rotate action, the log table data is optionally e-mailed to someone, and then the data is erased from the log table.

3.6.4 Email

Syntax: A valid e-mail address

When the query log is rotated (according to the schedule set), an e-mail message with an attached file (containing the previous log data) is sent to this address. Multiple addresses may be specified, separated by commas.

3.6.5 Result Order

Syntax: select Relevance or Date button

This determines the default ordering of search results. By default answers are ordered by rank (or relevance). Selecting "Date" makes search results ordered by date descending (newest first) by default. Search users may select the alternate ordering from this default.

3.6.6 Results Style

Syntax: choose from drop down list

This controls the style used for displaying individual answers to user queries. There are various styles from which to choose. The arrangement and amount of information varies in every style. In the administrative interface you may click the question mark (?) next to "Results Style" to see a sample of each of the available styles.

3.6.7 XSL File

Syntax: Browse local hard drive for a XSL file

This allows the use of a customized XSL file to format the output of a search. A default XSL style sheet is included with the Search Appliance. This option is used only if the `Results Style` of XSL Stylesheet is selected. The links below this option display the current XSL stylesheets.

3.6.8 Abstract Style

Syntax: choose from drop down list

This setting controls the short description or abstract that is generated for each search result. Choosing `Query` uses a snippet that matches the query. `Beginning` uses the start of the document's content. `Top` uses the top of the current page. `Description` uses the value of the `Description` meta tag.

3.6.9 Abstract Length

Syntax: enter number in text box

This determines the length in bytes of the document abstract.

3.6.10 Results per Page

Syntax: a whole number

This controls the number of results (answers) listed on each results page. When there are more than this many answers to a user's query the user will have to hit "next" to see more answers.

3.6.11 Results per Site

Syntax: an integer select box and Yes/No button

The **Max** setting controls the maximum results per site per page to display. For large profiles with many sites and many pages per site, limiting results per site can increase the variety of sites shown to the user, by avoiding "bunching" of results around a single site. It does reduce query speed however. When results are site-limited, the second and later results for a site are grouped together, indented under the first result for the site, and followed by a `More results for site` link.

The **Allow override** button controls whether the search user can override the profile's limit on the Advanced Search form.

3.6.12 Allow site: syntax

Syntax: a Yes/No button

This controls whether to allow the `site:host` query syntax in a search, to limit results to a single domain. It has no effect on the **From this domain** box on the Advanced Search form, which uses a separate variable (`sq`) instead of embedding in the query.

3.6.13 Results Width

Syntax: a whole number or a percentage valid for an HTML `<TABLE> WIDTH`

This controls the width of the `<TABLE>`s used in the search results. This may be a number indicating a fixed width or a number from 1 to 100 followed by a percent sign(%). This tells the user's web browser how wide to make the table.

3.6.14 Box Color

Syntax: a color name or number valid for HTML color specification

This controls the color of the “gray” informational boxes at the top and bottom of search results pages.

3.6.15 Display Thunderstone logo on results

Syntax: select Yes or No button

This controls the display of the Thunderstone logo on the search results page. The logo is displayed on the first and last page of a search.

3.6.16 Show Advanced

Syntax: select Yes or No button

This controls whether or not the Advanced Search button is displayed on the search form. If set to No then the button will be hidden, otherwise it will be displayed.

3.6.17 Font

Syntax: a font name valid for HTML `` specification

This specifies the font to use throughout the search interface.

3.6.18 Display Charset

Syntax: a standard IANA charset name

This sets the charset used to display search results in. The default if empty is the charset for Storage Charset under All Walk Settings. This charset should be a superset of US-ASCII (same 7-bit sequences), compatible with Top HTML, and translatable by the Search Appliance from Storage Charset.

A `<META HTTP-EQUIV=Content-Type>` tag in Top HTML will be updated automatically to reflect this charset. This update can be disabled by putting 2 or more spaces between `META` and `HTTP-EQUIV` in Top HTML.

Note that if the Display Charset differs from the Storage Charset, search results must be converted on-the-fly, potentially degrading performance slightly. Thus, if Display Charset is ever changed, it is recommended that Storage Charset be changed as well, and after the next rewalk (when all the database data is now in the new Storage Charset), Display Charset be change back to default (empty, which will still display in the new Storage Charset).

3.6.19 Top HTML and Bottom HTML

Syntax: HTML

This is static HTML to place at the beginning and ending of every search page respectively. It is useful for setting styles and displaying navigation menus and otherwise making the search pages look like the rest of your site.

Top and Bottom HTML when placed together should be exactly what is required to create a complete and valid HTML page. You can use your favorite HTML editor to create a page with a placeholder for the search form and results. Then cut and paste the section of HTML before the placeholder into the Top HTML and the section of HTML after the placeholder into the Bottom HTML.

If `$query` occurs within these fields, it will be replaced by the user's query.

3.6.20 Enable Sherlock

Syntax: select Yes or No button

This informs the search to include comment tags in the results page to allow Sherlock to process the list.

Sherlock is a metasearch tool for Macintosh computers.

3.6.21 Apply Appearance and Revert Appearance

Syntax: select checkbox

Changes made to the search settings are not normally immediately visible to end users. They may be tested using the "Test Search" menu item. This allows you to see the effects of your changes before committing to them.

Selecting `Apply Appearance` causes the settings currently shown on the form to be made live so that end users will see them. Once this is done, it is permanent, and you must edit the settings to get back the earlier appearance. There is no undo.

Selecting `Revert Appearance` causes the unapplied search settings to be discarded. The settings on the form are reset to those being used on the live search.

3.6.22 Top Best Bet Title

Syntax: text

This is the title text of best bets displayed above the search results. Common choices are “Best Bets” and “Suggested Links”. See *Using Best Bets* 4.14 for more details.

3.6.23 Right Best Bet Title

Syntax: text

The title text of best bets displayed to the right of search results. Common choices are “Best Bets” and “Suggested Links”. See *Using Best Bets* 4.14 for more details.

3.6.24 Top Best Bet Group

Syntax: choose group from drop-down list

This controls which group of best bets will be shown above the results. The group must already be created. See *Using Best Bets* 4.14 for more details.

3.6.25 Right Best Bet Group

Syntax: choose group from drop-down list

This controls which group of best bets will be shown to the right of the results. The group must already be created. See *Using Best Bets* 4.14 for more details.

3.6.26 Top Best Bet Box Color

Syntax: valid HTML color

This controls the color to be used for the background of the top best bet box. See *Using Best Bets* 4.14 for more details.

3.6.27 Right Best Bet Box Color

Syntax: valid HTML color

This controls the color to be used for the background of the right-side best bet box. See *Using Best Bets* 4.14 for more details.

3.6.28 Top Best Bet Border Style

Syntax: select from drop-down list

This controls the style of the top best bet box border. You can choose to have no border, a border around all the best bets, or an individual border around each result. See `Using Best Bets 4.14` for more details.

3.6.29 Right Best Bet Border Style

Syntax: select from drop-down list

This controls the style of the right-side best bet border. You can choose to have no border, a border around all the best bets, or an individual border around each result. See `Using Best Bets 4.14` for more details.

3.6.30 Right Best Bet Box Width

Syntax: enter number in text box

This controls the width of the best bet boxes shown to the right of the regular search results. See `Using Best Bets 4.14` for more details.

3.6.31 Authorization Method

The `Authorization Method` setting controls what Results Authorization method(s) are used by the Search Appliance when verifying user access to search result URLs. See the Results Authorization section (p. 63) for details. The possible settings are:

- `None`: No access verification; return all search results to all users. This is the default.
- `Forward login cookies`: The Search Appliance will forward login cookies from the user to the result URL. This is for custom HTML-form-based single-sign-on systems.
- `Basic/NTLM/file - prompt via form`: The Search Appliance will prompt the user for their credentials with a form, then send them to the result URL via HTTP Basic, NTLM or Windows/SMB file authentication.

3.6.32 Login Cookies

For the `Forward login cookies` Results Authorization method, one or more cookies must be named in the `Login Cookies` setting. No values are given, as they will be obtained automatically on a per-search basis from the user.

When a user conducts a search, if the named cookies are seen from the user's browser, the user is assumed to be logged in, and the cookies are forwarded to the results URLs for authorization. If the named cookies are not seen, the user is assumed not to have logged in yet, and is redirected to `Login URL` instead.

3.6.33 Login URL

For the `Forward login cookies Results Authorization` method, if none of the `Login Cookies` are seen at search time, the user is assumed not to have logged in yet, and will be redirected to this URL instead. The `Login URL` should be the URL to the site's form-based login page.

After logging in, the site's login page can be configured to re-redirect the user back to their original search if desired. The special token “%REFERER%”, if used in the `Login URL`, will be replaced with the URL back to the user's search. Thus, it could be assigned to a query-string variable in the `Login URL` so that the login page can redirect back to the search. Eg. with this value for the `Login URL`:

```
http://login.acme.com/login.asp?searchurl=%REFERER%
```

the Search Appliance would redirect the user to `http://login.acme.com/login.asp`, with the `searchurl` variable set to the Search Appliance search page (with query). The `login.asp` code could be modified to redirect the user back to the `searchurl` query variable after login.

3.6.34 Basic/NTLM/file Cookie Type

For the `Basic/NTLM/file - prompt via form Results Authorization` method, this setting controls what cookie type to use for the Search Appliance's copy of the user's credentials.

With `Basic/NTLM/file - prompt via form` set, when a user conducts a search for the first time, a form is presented (from the Search Appliance) asking for a user and password. The user/pass is sent back to the user as a cookie from the Search Appliance for use in future searches without having to re-prompt. The user/pass is also simultaneously used to validate search results via HTTP Basic/NTLM or Windows/SMB file access.

The `Basic/NTLM/file Cookie Type` setting controls whether this cookie from the Search Appliance should be `Persistent` (retained permanently so the user does not have to login again) or `Session` (discarded after browser closure for security).

Note that the `Basic/NTLM/file Cookie Type` cookie is distinct from the `Login Cookies`; they are used for different access methods. The former originates from the Search Appliance and is only ever sent to/from the user and the Search Appliance: non-cookie-based access methods are then used from the Search Appliance to the result URLs for actual authentication. `Login Cookies`, however, originate from a third-party form-based login system, and pass from the login server to the user to the Search Appliance to the result URLs.

3.6.35 Login Verification URL

For the `Basic/NTLM/file - prompt via form Results Authorization` method, the user is directly prompted for a login by the Search Appliance. Since authentication is handled by another server, when search results are denied access, the Search Appliance cannot know if the denial is URL-based (lack of access by the user), or login-based (mistyped/wrong password).

To differentiate the two and give users a chance to correct mistyped passwords, a `Login Verification URL` may be set. This should be a URL that *all* users have access to, but that is still protected (ie. anonymous users are denied). It should be an actual file (not a directory), preferably small (a few KB), and permanent (not likely to move, be renamed or have perms changed).

If `Login Verification URL` is set, the Search Appliance will verify a user's prompted-for login by accessing this page. Since all users have access to it, a denial is assumed to mean the login was incorrect, and the user will be re-prompted for their credentials. Without a `Login Verification URL` set, a mistyped password will result in no search results, but the user won't know if they do not have access to the results, or they merely mistyped their password.

3.6.36 Unauthorized Result Query

For all `Authorization Method` types of `Results Authorization`, it is assumed a protocol-level denial will be issued when the Search Appliance accesses URL(s) that a user does not have access too. Eg. for HTTP URLs, a 401 `Unauthorized` message should be issued.

However, some servers may only issue a human-readable denial message, but otherwise return an ok (eg. HTTP 200) protocol message. For such results the Search Appliance will assume the user has access, and will erroneously return the result.

To remedy this, `Unauthorized Result Query` may be set to a query that will match only denied pages (eg. "Access Denied"). The `Field/Type` box should be set to the query type (substring vs. REX) and field (raw HTML vs. formatted text) for the search. The `Query` field is set to the actual substring or REX query.

Note that this setting imposes an extra search load, as each search result fetch must now be a GET instead of a HEAD, as well as queried against. Thus, `Unauthorized Result Query` should only be set if absolutely necessary.

3.6.37 Max Docs to Auth-Check

This setting is the maximum number of raw (pre-auth-check) search result URLs to examine for authorized results, during results authorization. Decreasing this limit can speed up searches and reduce origin server load, at the cost of possibly truncated displayed results. Eg. noisy queries that match many overall documents on the server, but few of which are authorized for the search user, may use a lot of server resources, so reducing this limit may reduce that load.

The maximum value is -1 or blank (the default), for no limit: ie. continue until all results are checked, or `Successful Auth Result Limit` or `Total Auth Timeout` is reached.

3.6.38 Successful Auth Result Limit

This setting is the maximum number of authorized (displayable, post-auth-check) results to try to establish, during results authorization. Increasing this limit makes it more likely to get an exact hit count for a search (instead of a single page), at the expense of more search time and more origin server load.

The minimum (and default if empty) is the same as the `Results per Page` setting (p. 50), which produces a page of results the fastest. The maximum is -1 for no limit, ie. continue until all results are checked, or `Max Docs to Auth-Check` or `Total Auth Timeout` is reached.

3.6.39 Total Auth Timeout

This setting is the maximum total time in seconds to spend searching and authorizing results, during `Results Authorization`. The maximum setting value is -1 for no limit, ie. let `Search Timeout` (p. 60) cancel the search if reached. Any other negative value is relative to `Search Timeout`. Thus the default (if empty) of -5 means stop searching 5 seconds before `Search Timeout`, so that there are a few seconds left to send the results to the user.

3.6.40 Debug Results Authorization

Enabling this setting causes copious debugging information to be logged. It should only be enabled at the request of Tech Support for diagnosing Results Authorization problems.

3.6.41 Enable Spell Check

Syntax: select Yes or No button

This turns on the spell check option. With this option on, any search which produces no results displays a list of alternate-spelling queries, which will produce more results. If a query produces one result, the Search Appliance suggests other words similar in spelling to the words you entered. The suggestions are based on the actual walk database, so unusual spellings or terminology used on your site are picked up by the spell-checker. The number of suggestions varies, depending on the `Suggest Time Limit` and `Number of Suggestions` options. The default is on.

3.6.42 Suggest Time Limit

Syntax: choose from drop-down list

This controls the number of seconds the Search Appliance allows for spelling suggestions to be made. See also `Enable Spell Check` 3.6.41 for more information.

3.6.43 Number of Suggestions

Syntax: choose from drop-down list

This controls the number of spelling suggestions offered. See also `Enable Spell Check` 3.6.41 for more information.

3.6.44 Synonyms

Syntax: choose from drop-down list

This allows you to select a level of equivalence matching. You can limit results to specific matches, or you can allow synonymms and phrases. The values are described as follows:

`Disabled`: no phrase recognition and no synonyms (equivalences). Only searches for the the actual terms in a query. This is regardless of `~` usage.

`Phrase recognition only`: recognize query word groups that are known phrases and search for them as phrases.

`Phrases & Allow synonyms`: phrase recognition plus allows the tilde (`~`) operator to match synonyms on specific query terms

`Phrases & Use synonyms by default`: phrase recognition and matching synonyms on all query terms (tilde to turn off on specific terms).

See also `Using the Thesaurus` (section 4.3).

3.6.45 Main Thesaurus

Syntax: the symbolic name for the primary thesaurus

Here you can select a main thesaurus. A drop-down list allows you to select one of the thesauri that was defined in `Maintenance, Custom Thesaurus`.

See also `Using the Thesaurus` (section 4.3).

3.6.46 Secondary Thesaurus

Syntax: the symbolic name for the secondary thesaurus

Here you can select a secondary thesaurus. A drop-down list allows you to select one of the thesauri that was defined in `mMaintenance, Custom Thesaurus`.

See also `Using the Thesaurus` (section 4.3).

3.6.47 Allow the @ Operator

Syntax: select Yes or No button

Off by default. If on, allow use of the `@` (intersections) operator in queries. Queries with few or no intersections (eg. `@0`) may be slower, as they can generate a copious number of hits.

3.6.48 Allow Linear

Syntax: select Yes or No button

Off by default. If on, an all-linear query –one without any indexable “anchor” words– is allowed. A query like `“/money #million”`, where all the terms use unindexable pattern matchers (REX, NPM or XPM) is an example. Such a query requires a linear search of the entire table, and this can be very slow for a table of significant size.

If `alllinear` is off, all queries must have at least one term that can be resolved with the Metamorph index, and a Metamorph index must exist on the field. Under such circumstances, other unindexable terms in the query can generally be resolved quickly, if the “anchor” term limits the linear search to a tiny fraction of the table. The error message `“Query would require linear search”` may be generated by linear queries if this is off.

3.6.49 Allow NOT Logic

Syntax: select Yes or No button

On by default. If on, allows “NOT” logic (eg. the `-` operator) in a query.

3.6.50 Allow Post-Processing

Syntax: select Yes or No button

Off by default. If on, post-processing of queries is allowed when needed after an index lookup, eg. to resolve unindexable terms like REX expressions, or only partially indexable terms. If off, some queries are faster, but they may not be as accurate if they aren’t completely resolved. The error message `“Query would require post-processing”` may be generated by such queries if this is off.

3.6.51 Allow Wildcards

Syntax: select Yes or No button

On by default. If on, wildcards are allowed in queries. Wildcards can slow searches because potentially many words must be looked for.

3.6.52 Allow WITHIN Operators

Syntax: select Yes or No button

Off by default. If on, “within” operators (`w/`) are allowed. These generally require a post-process to resolve, and therefore they can slow searches. If off, the error message `“‘delimiters’ not allowed in query”` will be generated if the within operator is used in a query.

3.6.53 Resolve Phrase Noise Words

Syntax: select Yes or No button

Off by default. This indicates whether to exactly resolve the noise words in phrases. If on, a phrase such as “state of the art” will only match those exact words; however, this may require post-processing to resolve (potentially slower). If off, any word is permitted in place of the noise words, and no post-processing is needed; this is faster but potentially less accurate.

3.6.54 Keep Noise Words

Syntax: select Yes or No button

Off by default. This indicates whether to remove noise words from the query during query processing.

3.6.55 Search Timeout

Syntax: integer number of seconds

This is the maximum overall time to spend searching and sending results. Exceeding this limit, whether due to server load, network slowness, etc. will result in a “Timeout” message to the user. This helps prevent heavy load from overwhelming the server. The default (if empty) is 30 seconds. The maximum is -1 for no limit, which is strongly discouraged.

3.6.56 Fast Result Counts

Syntax: select Yes or No button

Off by default. Some complex queries involving categories or proximities closer than page can take more time to determine exact result hit counts. In some cases it may cause timeouts. Enabling this option will determine hit counts much faster, and using less CPU, in these cases at the expense of accuracy. The hit counts for complex queries will generally be overestimated (it will say there are more hits than there really are).

3.6.57 Proximity

Syntax: choose from drop-down list

Proximity gives the ability to locate answers with greater precision. The Search Appliance input form gives you several options to control the search proximity:

line All query terms must occur on the same line

sentence Query items must all reside within the same sentence

paragraph Within the same paragraph or text block

page All items must occur within same HTML document (the default)

A bar-graph display will be shown any time a ranking search was performed (eg. all searches except Show Parents).

3.6.58 Word Forms

Syntax: choose from drop-down list

The **Word forms** options give you control over how many variations of your query terms are sought in your search as follows:

Exact: Only exact matches are allowed. (the default)

Plural & possessives: Plural and possessive forms are found. (s, es, 's)

Any word forms: As many word forms as can be derived are located.

3.6.59 Word Ordering

Syntax: choose from drop-down list

Controls how important word order is for results ranking: hits with terms in the same order as the query are considered better. For example, if searching for “bear arms”, then the hit “arm bears”, while matching both terms, is probably not as good as an in-order match. The default weight is Medium (500).

3.6.60 Word Proximity

Syntax: choose from drop down list

Controls how important proximity of terms is for results ranking. The closer the hit’s terms are grouped together, the better the rank. The default weight is 500.

3.6.61 Database Frequency

Syntax: choose from drop down list

Controls how important frequency in the table is for results ranking. The more a term occurs in the table being searched, the *worse* its rank. Terms that occur in many documents are usually less relevant than rare terms. For example, in a web-walk database the word “HTML” is likely to occur in most documents: it thus has little use in finding a specific document. The default weight is 500.

3.6.62 Document Frequency

Syntax: choose from drop down list

Controls how important frequency in document is for results ranking. The more occurrences of a term in a document, the better its rank, up to a point. The default weight is 500.

3.6.63 Position in Text

Syntax: choose from drop down list

Controls how important closeness to document start is for results ranking. Hits closer to the top of the document are considered better. The default weight is 500.

3.6.64 Ranked Rows

Syntax: number

The maximum number of rows that can be scrolled to when returning ranked results. This can be set to 0 for all matching rows, or to any other number. The lower the number the better the performance, however users won't be able to scroll through as many results. The default is 200.

3.6.65 XML Export Variables

Syntax: names separated by newlines

XML Export Variables is a list of variables, one per line, that are to be displayed and propagated through XML search results. For example, if `cbtGroup` is specified in XML Export Variables, and the search query includes `... \&cbtGroup=user\&cbtGroup=backup...`, then the following block will appear in XML output, after all the `<Result>` tags:

```
<exportVar>
  <variable name='`cbtGroup`'>user</variable>
  <variable name='`cbtGroup`'>backup</variable>
</exportVar>
```

This setting only applies if `Results Style` is set to `XSL Stylesheet` and `Allow XML` is asserted.

3.6.66 File Url Format

Syntax: choose from drop down list

Controls how file urls are formatted. The `Max Compatibility` setting will format them as `file:///server/share` which both Internet Explorer and Mozilla-based browsers such as Firefox and SeaMonkey support, while the Internet Explorer only setting will format them as `file://server/share`.

Note that for Mozilla based browsers you will also need to enable permission for HTML pages to open files by creating a `user.js` in your profiles directory (where `prefs.js` is) that contains (note: lines wrapped to fit the printed page):

```
user_pref("capability.policy.policynames", "localfilelinks");
user_pref("capability.policy.localfilelinks.sites",
          "http://APPLIANCE");
user_pref("capability.policy.localfilelinks.checkloaduri.enabled",
          "allAccess");
```

where `APPLIANCE` is the name or IP address of the search results as seen in the browser address bar.

3.6.67 Visible

This controls whether this profile is visible to other appliances (or even the same one) for use in a meta search. Any profile that is to be used as a part of a meta search must have the `Visible` flag set to `Y`.

3.7 Results Authorization

Results Authorization allows restriction of search results to authorized users only, on a per-URL basis. Only users with access to a given URL will ever see that URL in a result list, instead of all users seeing all matches (and potentially being denied access to results already shown).

Access to a URL, as well as the namespace of users, is determined by the URL's origin server, not the Search Appliance, so no reconfiguration of users or access is needed – the pre-existing server access controls are just forwarded by the Search Appliance. And since access is determined on a per-result, not per-search, basis, a single profile can serve a multitude of users with any combination of whole/partial access to the underlying data.

Results Authorization works at search time by accessing each potential search result URL with the user's credentials. Only URLs authorized to that user are then shown in search results. The authentication method(s) used will depend on the existing system(s) already used by the indexed URLs. Various schemes are supported:

- **None:** No access verification; return all search results to all users. This is the default.
- **Cookie-based:** Custom HTML-form-based single-sign-on systems. Users first login on a web server (not a Windows workstation login), which then sends an access cookie to the user's browser. This cookie is automatically returned to the server when accessing future pages, and grants the user access.
- **Basic:** HTTP Basic authentication, for web servers.
- **NTLM:** Windows NTLM authentication, for web servers.
- **SMB/Windows:** SMB for Windows file servers.

For cookie-based systems, the Search Appliance will merely forward the cookies the user has already received from the site login page. For all others (Basic/NTLM/SMB), the Search Appliance must prompt for the user and password directly, as they are needed to verify result URLs. In the latter case, credentials will then be stored in a cookie by the Search Appliance so that future searches do not need to re-prompt for a login.

3.7.1 Results Authorization Crawl Settings

The Search Appliance itself needs read access to the entire set of URLs in order to build a search index. Therefore, before walking a protected data set for Results Authorization, it may be necessary to fill out the `Login Info` setting (p. 42) under `All Walk Settings` with a full-access admin type account, so that the Search Appliance can crawl the data.

Or it may be necessary to fill out a `Primer URL` (p. 42) containing login info to submit to a site's login form, so that the Search Appliance can obtain the login cookies needed for access to the rest of the site.

3.7.2 Results Authorization Search Settings

After a successful crawl, Results Authorization is configured with the `Results Authorization Options` group on the `Search Settings` page. The primary setting is `Authorization Method` (p. 54), which is determined by the authentication system(s) in use by the indexed URLs. If cookie-based, this is set to `Forward login cookies`; for all other systems, it is set to `Basic/NTLM/file - Prompt via form`. Most of the remaining settings depend on which method was selected; see the `Authorization Method` setting (p. 54) for details.

There are also a few resource/tuning settings, such as `Max Docs to Auth-Check`, `Successful Auth Result Limit`, `Total Auth Timeout`, and `Debug Results Authorization`, which are not required, but merely fine-tune the results.

3.8 Meta Search - Search multiple profiles as one

Meta search allows you to search multiple profiles simultaneously and merge and display the results as if it was one big profile. The meta search can search and combine profiles from multiple appliances.

3.8.1 Profile Creation

Create a Meta Search profile using the normal profile creation page but select as a copy of `-Meta Search Defaults-` instead of `-Defaults-` or some other profile.

3.8.2 Walk Settings

`Walk Settings` is somewhat of a misnomer for a meta search profile since it doesn't do any walking of its own. On this page you list the host(s) and profile(s) to search and merge when this profile is accessed.

For each profile you want included in the search you list the profile's name in the `Profile Name` column and the host name or IP of the appliance where that profile resides in the `Host IP or Name` column. You may use DNS resolvable names or IP addresses in the host column. IP addresses are slightly more efficient because they don't require DNS lookup. But names are more flexible. Only the DNS, not a bunch of profile settings, has to change when machines get replaced or renumbered.

Profiles on the same appliance will be searched serially (one after the other). Profiles on different appliances will be searched in parallel (at the same time). "Sameness" of appliance is determined by the `Host IP or Name` setting so using different names or a name and an IP for the same physical appliance will cause that appliance to be searched in parallel with itself.

The `Visible` setting controls whether this profile is usable by other appliances. You may have nested meta searches.

All appliances being accessed by a meta search must have the IP address of the meta search appliance listed in the `Cluster Members` setting under `Maintenance->System Wide Settings`

3.8.3 Search Settings

The appearance options control the appearance of the meta search results pages. Currently the `Results Authorization` and query options of the meta profile do not apply. Those of the backend profiles being queried apply. Login information is not propagated to the backend profiles so result authorization when using meta search is not currently possible, though it is planned for a future release.

When using best bets the meta search profile must have the same group names as the backend profiles. Any best bets from the backends that have group names that are not defined in the meta profile will not be shown.

Query logging of the meta search and the backends are independent of each other. The meta search will respect its own query logging setting as will each of the backend profiles. So it's possible to have multiple logs for the same query if both the meta search and the backend have query logging turned on.

3.9 Access Control

Access Control allows different administrative users to be given different levels of access to the Search Appliance; normally, with access control off (the default) all users have access to all administrative functions. Access Control can only be enabled or disabled by the `admin` user, on the `Maintenance` page.

3.9.1 User Groups

User groups allow easier access control maintenance, as users with similar permissions can be administered together once rather than separately several times. The special group `Everyone` always exists and cannot be edited; it always contains all users as a convenience.

User groups may contain other groups as well as users, allowing complex hierarchies to be created if needed. Permissions for a user are affected by all groups a user is directly or indirectly a member of. For example, if user Amy is in group `Programmers`, and group `Programmers` is in group `IT`, then Amy is

also indirectly a member of IT, and her permissions are affected by those granted to not only herself and Programmers but IT as well.

3.9.2 Object hierarchy

Each administrative action that can be access-controlled (eg. editing walk settings, creating accounts) can be thought of as an object. Some actions are broader than others and can be thought of as a superset, eg. editing *all* profiles is a superset of editing a *specific* profile. Thus, access control objects are arranged in a tree-like hierarchy, where each object has a parent object, and can inherit permissions from it. This makes setting privileges on a logical group of objects (eg. all profiles) easier, as only one object may need to be changed (the parent). Also, when new child members (eg. new profiles) are created, they will inherit the same privileges automatically. The access control object hierarchy in the Search Appliance is as follows:

/	Global root object
Users/	User accounts
admin	admin user
...	Other users
Groups/	User groups
Profiles/	Profiles
default	default profile
...	other profiles
Settings/	Profile settings
Maintenance/	Maintenance page
Info/	
Updates/	
Logs/	
Settings/	
System Wide	
ACLs	
Thesaurus	
Save, Restore	
Mounts	
System/	
RAID	

Note that these “files” do not really exist: the objects are merely symbols representing actions that can be access-controlled.

3.9.3 Access Control Lists

An object may have an Access Control List (ACL) associated with it. ACLs determine what rights (Read/Write/Delete/Change perms) users have on objects. Each object’s ACL contains one or more Access Control Entries (ACEs). An ACE identifies a trustee (a user or group), a set of rights, and whether those rights are allowed or denied the trustee on that object. In addition to the ACL explicitly set on an object, rights may be inherited from parent objects’ ACLs, as mentioned above.

3.9.4 Determining Effective Rights

The effective rights a specific user has on an object – what the user can actually do with the object – are determined by examining ACEs in a specific order. The first ACE that matches both the user and the desired access right determines whether the user has that right on the object. An ACE matches the user if it specifies the user or any group the user is directly or indirectly a member of. An ACE matches the desired right if the right is listed in the ACE.

ACEs are examined in the following order¹:

1. ACEs explicitly set on the object
2. ACEs explicitly set on the object's parent
3. ACEs explicitly set on the object's further ancestors, nearest ancestor first

At each object, ACEs are checked in ACL order (the order displayed for an object on the Access Control page). If no matching ACE is found after all levels are examined (back to the root or Global ACE), access is allowed by default (this is for back-compatibility with non-ACL mode).

3.9.5 Required Rights for Admin Actions

Certain ACL rights are required for certain administrative actions to be performed. In order to maximize rights-configuration flexibility, some actions require rights on multiple objects. For example, editing settings on a profile requires rights not only on the profile, but also on the setting itself. Note in the object hierarchy (p. 66) that profiles and settings are two “sibling” branches, rather than settings being replicated as descendants of every profile. Thus, profiles and settings can be thought of as a two-dimensional grid for permissions, and a user's rights can be tailored across that grid: access to one setting across all profiles, access to all settings one profile only, etc.

The rights needed for specific actions are listed below. If a user does not have the required rights for an action, either a red `Access denied` message will be displayed, or (if access still granted to other parts) the affected object may simply not appear (read access denied), or may appear grayed out (write access denied). For more information and some example permission schemes, see the Using Access Control section, p. 94.

Walk and Search Settings

For settings under Basic, All Walk, and Search Settings, a user must have read access to the profile as well as read access to the specific setting in order to see the setting. Write access to the profile, and write and delete access to the setting, is needed in order to modify a setting. (Delete is needed to clear a setting, which may not be apparent from the form.) Note that some settings are grouped on a line, such as the Enterprise setting: permissions can be granted to the group as a whole (Enterprise), or only specific settings in the group (Enterprise - Yes or Enterprise - Domain). If a user has no read

¹In versions 5.3.0 and earlier, deny ACEs were always required to be before allow ACEs for an object.

access to a setting, it will not be displayed on the page. If a user has no write access to a setting, it will be disabled (grayed out and not modifiable).

Starting and stopping a walk

Write access to the profile and write access to the `Walk now` setting is required to start a walk. Write access to the profile and write access to the `Stop walk` setting is required to stop a walk.

Best Bets

Write access to the profile and write access to the `Best Bet Groups` setting is needed to modify the Best Bet Groups for a profile, or to modify Best Bet words for a specific URL (under List/Edit URLs). Note that this is distinct from editing Best Bet *search* settings (eg. `Top Best Bet Title`), which only affect search, not the walk itself.

List/Edit URLs

Write access to the profile and write access to the `Link report` setting is needed to modify URLs in the database, including using the `Update Soon` link. Read access to both is needed to view URLs.

Walk Status

Read access to the profile and read access to the `Walk status` setting is needed to view Walk Status.

Query Log

Read access to the profile and read access to the `Query log` setting is needed to view the Query Log.

Profiles

Read access to the profile and read access to the desired setting(s) are needed to view the given setting. Write access to both is needed to modify a setting. Delete access to the profile is needed to delete the profile. Write access to `All Profiles` (the parent of profiles) is needed to create a new profile.

Accounts

Write access to `All Users` is needed to create a new user. Write access to the user is needed to change the password for a user. Delete access to the user is needed to delete a user.

User Groups

Write access to `All Groups` is needed to create a new group. Write access to the group, as well as write access to each member being added or removed, is needed to add or remove members to or from a group (except where the group is only indirectly being modified due to a member itself being deleted). Delete access to the group is needed to delete a group.

Access Control

Change-perms access to an object is needed in order to create, edit or delete an ACE on the object.

Maintenance

Read access to `Info` under `Maintenance` is needed to read the `Information` links. Write access to `Updates` under `Maintenance` is needed to install or upgrade software. Read access to `Logs` under `Maintenance` is needed to read the `Logs` links. Write access to `Settings` under `Maintenance` is needed to modify `Settings`, *with the exception* of `Enable` or `Disable Access Control Lists`: these can *only* and *always* be performed by the admin user, *regardless* of ACLs (eg. for emergency reset). **Note:** giving a user write perms on `Settings`, directly or indirectly, can allow them to override anything on the system, eg. via externally-modified save and restore settings. Also, note that a user with physical access to the machine could overwrite settings, eg. re-install the software.

3.10 Running the Search Interface

See section 4.1, p. 83.

3.11 Maintenance

The Maintenance menu has the following structure. Each item is described in the pages that follow.

- Information
 - Display disk space
 - System Information
 - Thunderstone Information
 - Tech Support Information
- Install/Upgrade
 - Setup/edit update preferences
 - Check for updates
 - Install from CD
- Logs
 - Manage logs (View, Delete, Rotate, Send)
- Search Appliance Settings
 - System Wide Settings
 - View/Edit Access Control Lists
 - Enable or Disable Access Control Lists
 - Custom Thesaurus
 - Manage SSL/HTTPS Server Certificates
 - Save Search Appliance settings
 - Restore Search Appliance settings
 - Network Filesystems & Shares
 - DBWalker Settings / Status
- Appliance system access
 - RAID Array Management
 - Webmin Interface

3.11.1 Information

The Information group provides links to a variety of information useful for monitoring the system and performing maintenance.

Display disk space

This page provides disk space information (used and available). The information is presented at the bottom of the page.

System information

This page provides system information including: network IP addresses, MAC address, kernel version, load, and time.

Thunderstone information

This page provides Thunderstone software version numbers, the Appliance serial number, and Thunderstone contact information.

Tech support information

This page provides a convenient means of communicating your system information to Thunderstone technical support. Click the Email link to automatically send the information.

3.11.2 Install/Upgrade

The `Install/Upgrade` provides links to pages for installing and upgrading software.

Setup/edit update preferences

This allows you to configure the system to perform updates automatically. There are three steps to performing an upgrade, and you can select how many steps are performed automatically. The steps are described as follows:

- Discover Updates. Discovers whether software that is newer than what is installed is available.
- Download. Obtain newer software from Thunderstone through the Internet.
- Install. Install the downloaded software.

Check for updates

This allows you to manually initiate a check for software updates. It provides a list of available updates, allows you to select which updates to download, and allows you to manually initiate the installation of the downloads.

Refer to `Getting Software Updates 4.4` for the procedure to manually perform updates.

Install from CD

This allows you to install the software from a CD.

3.11.3 Logs

The logs provide detailed information about operational events of the database and the system.

Manage logs

This allows you to view, delete, rotate, and email the appliance logs. It lists every file in the log partition and allows you to manipulate each log individually or many at once. Each item has a check box for selection for mass operations, the date and time of last addition, the size, a link to see the most recent bit of that log, and a list of processes currently using that log. Clicking a column header will sort the list by that column.

There may be multiple versions of each log. The version with no numeric extension (.1, .2, etc.) in the filename is the current log. Those with numeric extensions are older logs. Extension .1 is the most recent old log, .2 is the second most recent, etc.. Logs are automatically rotated once a month or once a day if they exceed 10MB. The need for rotation is checked once a day around 4am. To force a rotation check you can click the `Rotate Logs` button at the top of the manage logs page. If you see a log file with numeric extension that also has process numbers listed in the `In Use By` column you'll probably need to reboot the appliance to free up that log file. That should be a rare occurrence, but has been known to happen.

The log listing is divided into sections. The first, unnamed, section is the system level logs. They contain information about the core operating system of the appliance. That's where hardware, network, and similar events and problems are logged.

The apache section contains the usage logs for the apache webserver which is used for HTTPS access to the appliance, if enabled.

The txis section contains logs related to Taxis, the relational database server that is a major technical component of the Search Appliance. These logs provide detailed information about operational events of Taxis.

The webmin section contains usage logs for the Webmin system management interface.

At the bottom of the page is a form which allows you to perform actions on the selected log(s). You may View, Delete, or Send any of the logs. For deleting you will be asked to confirm the deletion before it's carried out so it should be difficult to delete a log by accident. Deleting logs listed as in use may require rebooting to reclaim their disk space.

For viewing and sending you can choose how many lines of the log(s) to see and whether to see the newest lines first (reverse chronological order) or natural order with the oldest lines first.

For sending the logs to Thunderstone technical support you should fill in your email address and a ticket number given to you by Thunderstone. Sending the logs via email assumes that your appliance is configured to send mail to the internet. It is by default in the simple case but if you have outbound SMTP blocked by your firewall or need to use a mail relay you'll need to configure sendmail using the Webmin interface. An alternative to having the appliance email the log is to instead view the log then use your browser's "save as html" feature to save the raw html of the log view page to disk. Then attach that file to your email to Thunderstone tech support.

3.11.4 Search Appliance Settings

This area is for settings that affect the search appliance as a whole and/or may be shared by multiple walk profiles.

System Wide Settings

Home Page

By default when the appliance is directly accessed, as in `http://appliance_ip` it will present a page that allows selection of the admin or search interface. This option allows you to relace that page with any html you devise. The html you upload should refer to images and such using fully qualified URLs because they can not be uploaded to the appliance for use in relative URLs.

Checking `Default` will revert the appliance home page to it's factory behavior.

Enter At Search, Default Profile

By default users accessing the appliance using no particular URL will be given a choice of admin or search. Enabling this option removes that choice and enters at the search for the profile named in the Default Profile setting.

Favicon.ico

The appliance comes with no `favicon.ico` file. If you wish users' browsers to display your company's favicon when they are accessing the appliance you'll need to upload that icon. If you no longer wish to have a favicon check `Delete`.

Cluster Members

This field defines the machine(s) and/or network(s) that constitute a cluster of appliances. If you have more than one appliance all of their IPs or a network prefix and wildcard (such as `10.10.10.*`) should be specified here. All machines matching these IPs will be allowed full access to appliance internals without verification. This allows for replication and meta searching.

Enable HTTPS Admin

This enables the appliance's web based admin to accessed via HTTPS in addition to or instead of HTTP. Turn this on to enable encrypted communications. Then access the admin interface using `https` in the URL instead of `http`.

Require HTTPS Admin

Use this option so that the admin interface is only available via HTTPS and not HTTP. If you use this you should also turn on `Enable HTTPS Admin`. If you turn this option on while accessing via HTTP you will have to manually change your admin URL to use HTTPS instead.

Admin Access IPs

This controls what IP addresses are allowed to access the admin interface. You may specify one or more individual IP addresses or networks. Networks may be specified with either `address:netmask` or `address/prefixlen` syntaxes. Place each entry on a line by itself. Blank means no IP restriction, the admin interface may be accessed from any IP.

Example. If you have a local class C network of 10.10.1.0 as well as one public IP such as 198.49.220.1 you want to have admin access you would use

```
10.10.1.0/24
198.49.220.1
```

or

```
10.10.1.0:255.255.255.0
198.49.220.1
```

WARNING: Be careful when setting this option so that you don't block yourself from being able to admin. With improper settings the appliance could refuse admin from any IP you have access to. If that occurred you would have to reset the appliance to factory state (losing all crawl data) to regain access.

Enable SNMP service

This enables the SNMP server on the appliance. With this enabled you can use SNMP monitoring tools to monitor the condition of the appliance.

A few items of particular interest might be

What	OID (Object Identifier)
Disk space	.1.3.6.1.4.1.2021.9
System load	.1.3.6.1.4.1.2021.10
Critical processes	.1.3.6.1.4.1.2021.2

SNMP Community Name

This is the community name used to access the SNMP information. We suggest using something unique to your organization rather than "public".

SNMP Location Value

This is pretty much anything you want. It has no significance except as a designator for you to identify where or what the appliance is.

SNMP Contact Value

This is pretty much anything you want. It would normally contain some contact information for the admins of the appliance.

SNMP Access IPs

This controls what IP addresses are allowed to access the SNMP interface. You may specify one or more individual IP addresses or networks. Networks may be specified with either address:netmask or address/prefixlen syntaxes. Place each entry on a line by itself. Blank means no IP restriction, the SNMP interface may be accessed from any IP.

Example. If you have a local class C network of 10.10.1.0 as well as one public IP such as 198.49.220.1 you want to have SNMP access you would use

```
10.10.1.0/24
198.49.220.1
```

or

```
10.10.1.0:255.255.255.0
198.49.220.1
```

Debug Mode

This enables special debug functionality for crawling, searching and/or administration functions. It should be enabled *only* at the request of Thunderstone tech support, as debug mode can increase CPU and disk space (log) usage. Debug mode is sometimes needed by tech support to diagnose a particular problem. It enables an additional set of alternate URLs for debugging:

Normal URL	Debug URL
/taxis/dowalk/...	/taxis/dowalk-debug/...
/taxis/search/...	/taxis/search-debug/...

The mode can be set to one of 3 settings:

- **Off** This is the default setting: debugging is off, and debug URLs are disabled.
- **Alternate** Enables debugging for ...-debug URLs only; normal URLs have debugging off. This allows debug testing without disturbing live search.
- **Live** Enables debugging for both ...-debug and normal URLs. This allows debugging of live search, when the situation does not permit changing the URL (eg. an external URL fetch to the Search Appliance that cannot be altered).

When debugging is enabled, a banner message to that effect is printed at the top of every admin screen, to remind administrators that it is in effect (and should be disabled when debugging is finished).

Enable/Disable Access Control Lists, View/Edit Access Control Lists

This option turns on/off ACLs for accessing the appliance. The default permission scheme for managing the appliance is very basic and all accounts have full admin privileges. ACL's allow very fine grained control over which administrators can access which features and settings.

See section 3.9 for details about access control lists.

Custom Thesaurus

This area allows you to upload one or more custom thesauri (synonym lists) for use by search profiles. An uploaded thesaurus is compiled and kept on the appliance. There is no way to download a thesaurus once uploaded so it's a good idea to keep a copy around in case you want to make modifications later on.

Each thesaurus may be used by zero or more profiles and should not be deleted if it is in use by a profile. Search options that affect the use of these thesauri are `Synonyms(3.6.44)`, `Main Thesaurus(3.6.45)`, and `Secondary Thesaurus(3.6.46)`.

See section 4.3 for further details.

Save Search Appliance Settings

This allows you to save all of the current profile and most of the system settings from the appliance to an XML file on your local workstation. (Mounted filesystems and IP configurations are not currently saved.) This file can be used to aid in cloning appliances for a cluster and as a backup in the event the appliance needs to be restored from scratch.

Saving to floppy on the appliance or to a file location that the appliance has write perms on is not generally convenient. We recommend using the third option on this screen for saving to your local workstation. Right click the `here` link and select `Save Link As`, or similar, option from the browser menu that pops up.

Restore Search Appliance Settings

Use this option to restore settings that you've previously captured using `Save Search Appliance settings`.

Network Filesystems & Shares

Use this interface to mount remote file server(s) to the appliance so that it may be indexed into one or more walk profiles.

All created mounts are permanent until manually removed. They will be remounted upon reboot of the appliance.

To mount a remote filesystem or share select the type from the drop-down list and click `Add`.

NFS filesystems - Unix/Linux/etc. servers

Server: Enter the hostname of the server. CaSe does not matter. (eg: `nas1.mycompany.com`)

Directory: Enter the full path of the directory to mount as it is exported from the server. (eg: `/documents/internal`)

Reliability: Select `Hard` or `Soft`. `Hard` will cause the appliance to keep retrying the same file forever in the event of an error reaching the server. `Soft` will allow files to fail if the NFS server can not be reached.

NFS Version: The highest NFS version to use. Leave this at 3 unless you have problems with old NFS servers.

SMB - Windows share

Server: Enter the hostname of the server. CaSe does not matter. (eg: `nas1.mycompany.com`)

Share: Enter the name of the share as exported by the server. (eg: `internal`)

Login Name: Enter the login name for the account to use to access the files on this share. This should be a user that has permission to read all the files that need to be indexed.

Login Password: Enter the password for the selected login name.

Server IP: Rarely used. If the "Name" of the computer is different than it's DNS name it may reject mount requests to the "wrong" name. In that case enter whatever name makes the server happy into the `Server` field and enter the machine's IP address into this field.

NOTE: When using Windows 2003 server you may need to change a setting on the server to allow mounting from the appliance. If the share won't mount try setting
control panel->admin tools->domain security policy->security settings->
local policy->security options->
Microsoft network server: Digitally sign communications (always) to disabled.

Current mount list

Under the Add form is the list of currently mounts and their status. Each mount has a `Remove` link to unmount the filesystem and remove it from the list. The options for each mount may be clicked to examine or modify the options and remount the filesystem.

If an entry shows as "unmounted" there is a problem with the settings and it is not able to be mounted as is. If it was a transient problem with the server click the options then click "Save changes" without making any changes to retry the mount.

Under the options for each mount is also an example of the minimum `Base URL` you would enter into a profile to index the files on that filesystem.

The `Technical Info` link shows some internal details about the mounts that may be helpful to tech support if you have problems.

Note: This feature appeared in scripts version 5.4.11. Prior to that webmin was the only way to manage remote mounts.

3.11.5 Appliance system access

This area provides access to the management of the Appliance operating system and hardware which is not directly related to indexing and searching.

RAID Array Management

Note: This area only applies to larger models (such as 3000) that include multiple hot-swap disks in a RAID-5 configuration.

Note: The pages in this area may load somewhat slowly as they collect information from the RAID controller.

Overview

Little to no maintenance of the RAID array is required. In the event of a disk failure the hot-spare will automatically take over and the array will be automatically rebuilt. The rebuild process takes several hours. After the array is rebuilt the failed disk will have to be manually removed (no shutdown required) and replaced with a same size or larger disk of the same type. After the disk is replaced it needs to be Added to the array.

Details

The RAID Status page displays a summary of the RAID's state. It's an abbreviated form of the information on the RAID Management page to provide a quick Good/Bad check.

The RAID Management page lists information about the overall RAID array as well as each of the hard disks in the system. Each item starts with a Status and is color-coded to indicate it's state. Green is good, red is problem, blue is hot-spare disk, light blue is unused disk, yellow is verifying/testing.

The first line of the Storage table contains information about the overall array with the Use column set to Array. The remaining lines are individual disks, either Member, Spare, or None. Member disks are part of the RAID array. Spare disks are hot-spares that will take over for a failed member disk.

Each item in the Storage table has associated actions that may be taken.

Rebuild An array that is in a non-optimal state may be forced into a rebuild to become optimal again.

Verify Verifies the integrity of the parity information for the array. This is not generally needed as the array is automatically verified periodically as controlled by the hardware BIOS.

Fail Forces an individual disk into a failed state so that it may be replaced. This is not generally needed as failures will be automatically detected.

Remove This removes a hot-spare disk from the array. It then becomes an unassociated disk with a Use of None. All arrays should have a hot spare.

Add This adds an unassociated disk with a Use of None to the array as a hot-spare.

The first number in the Disk:Addr column (everything up to the :) is the disk number which corresponds to labels on the front panel of the appliance.

Rebuild/Verify Rate

The Rebuild/Verify Rate is how aggressively the RAID will rebuild. A higher rate will rebuild a partially failed array more quickly so that it's in a non-fault-tolerant state for the shortest possible time. The downside of the higher rate is that operations that use the disk such as walks and searches will see slower performance.

Controller

This table shows various model and version information about the RAID controller.

Command

The Command input box should not be used except at the request of Thunderstone technical support. It is for issuing arbitrary commands to the RAID controller. Putting the wrong thing in this field could irrevocably damage the RAID array and render your appliance completely unusable! If the "Ok to run this command" checkbox is not also checked anything in the command input will be ignore.

Perform

The Perform button at the bottom will perform all of the actions selected on the form. You must also set "Are you sure you want to perform these actions?" to "Yes" or the actions will not be performed.

Front panel

This provides a rough approximation to the physical front panel of the appliance. It shows the drive arrangement to aid in locating the proper disk when performing maintenance.

Webmin Interface

This area has its own login and allows for control of various low-level system settings. The login is `admin` using the same password as the `admin` account in the normal interface. If the password gets out of sync somehow it may be reset by setting the admin password from the Accounts area (3.3.12).

- **Remote mounts** Mount/attach remote NFS filesystems and Windows shares so that they may be indexed using direct `file://` urls. This method remains available but the newer `Network Filesystems & Shares` (3.11.4) method on the Maintenance page is simpler. See also `Indexing File Servers` (4.16).
- **Network** Configure the IP address, DNS servers, and routing.
- **Firewall** Restrict access by IP.
- **Clock** Synchronize the appliance to your local time.
- **Email delivery** Configure how to send email for walk notifications etc.
- **Shutdown** Shut the system down cleanly and power off.

Remote mounts using Webmin

Although the newer Network Filesystems & Shares (3.11.4) is preferred it may rarely happen that you need more options that are available in the webmin interface. In that case use these instructions to mount file servers using webmin.

To mount a file server you need to go to the Maintenance page, then the Webmin Interface, login as admin, and choose Disk and Network Filesystems. You can then Add mount of either NFS (Unix) or Windows Filesystem. When mounting a Windows filesystem the format for the “Mounted As” field should be /SERVER/SHARE (note: single forward slashes) and the “Server Name” and “Share Name” fields should also be filled in (without slashes). The server name in the “Mounted As” and “Server Name” fields should be all lowercase.

For Windows shares you can also specify a user name and password to access the fileserver, and the appliance will walk with those permissions. Use either a username local to the machine owning the share or use a domain login that the server understands. Domain logins have the form DOMAIN\USER.

Note: If mounting a Windows 2003 share fails with a message such as “18048: session setup failed: ERRDOS - ERRnoaccess (Access denied.)”, then on the Windows 2003 server you may need to change the setting Control Panel / Administrative Tools / Local Security Policy / Microsoft network server: Digitally sign communications (always) to be disabled, or perhaps the Domain Member: Digitally encrypt or sign secure channel data (always) setting may need to be disabled.

Generally you will mount the filesystem to save and mount at boot, so the files are always available, and you can set it to read only if desired, which will make sure that the appliance does not write anything to the fileserver.

When walking the Base Url would be file://SERVER/SHARE/. The trailing slash is required for directories. The case of SERVER and SHARE should agree with the “Mounted As” setting.

Example: Windows Server

Windows Networking Filesystem Mount Details

Mounted As : /corpserv/Repository
Server Name: corpserv
Share Name : repository

Advanced Mount Options

Login Name : AllowedUserName
Login Password: *****
Read-only? : Yes

Basic Walk Settings

Base URL : file://corpserv/Repository/Documents/

Example: NFS Server

Network Filesystem Mount Details

Mounted As : /corpserv/Repository
NFS Hostname: corpserv
NFS Directory: /Repository

Advanced Mount Options

Read-only? : Yes
Allow user interrupt?: Yes

Basic Walk Settings

Base URL : file://corpserv/Repository/Documents/

Chapter 4

Procedures and Examples

4.1 Searching your Index

Search the pages you have indexed by entering the following URL into your Web browser:

```
http://www.mysite.com/texis/search/
```

The above is a virtual path comprised of 2 parts. “.../texis” is the Taxis Web Script interpreter and “/search” is the path to the search script relative to your installation’s `ScriptRoot` (`/usr/local/morph3/taxis/scripts`).

The URL given above will search the live database specified in the default profile called “default”. If that profile is not found it will try to search the default walk database.

You may specify an alternate profile by including its name in the URL.

```
.../search/?pr=MYPROFILE
```

Where `MYPROFILE` is the name of the profile you wish to use. The search will use the live database specified by that profile.

You may also specify a database to search instead of a profile.

```
.../search/?db=DATABASE
```

Where `DATABASE` is the name of the database you wish to use. This would generally be the live database for a given profile which may be found as the first item listed on the administrative interface’s `Walk Settings` page. Databases used this way must exist under the `taxis` subdirectory of the installation directory. What you specify for `DATABASE` is only the portion of the path and name under the `taxis` directory. For example, to search the database `/usr/local/morph3/taxis/myprofile/db2` you would use:

```
.../search/?db=myprofile/db2
```

When using a database instead of a profile, the look and feel settings will be those that were live when the walk of that database was performed. The profile will not be consulted for more recent changes. A benefit of not consulting the profile, however, is some increased search speed, which may be useful on a very heavily searched system. A disadvantage of specifying the database is that it will no longer be correct if a new walk is performed.

To get help on constructing queries click on the **Advanced** button of the search form. On the advanced search form you will find hyperlinks into the search help, which is also included in this manual in section 6.

To place the search form onto your existing web page(s) call up the **Live Search** from the administrative interface main menu (or the URL you determined from the above). This will bring up the search form. Use your web browser's view page source option (MSIE: **TopMenu->View->Source**, Netscape: **TopMenu->View->Page Source**) to get the source of the page. Cut everything between and including the `<FORM>` and `</FORM>` tags. That form may then be pasted into the web page(s) of your choice. You may also rearrange the look of the form as long as the variables are still present. If you have categories there will be a `cq` select list in the form. You may leave this out if you always want to search everything. Or you may make it a hidden variable with a fixed value if you always want to search the same section.

4.2 Similarity Searching

The search script has a feature called "Find Similar" which allows a user to click on a search result record to find more pages within the database similar to that one. This feature may also be accessed from any web page by placing the appropriate URL on it. You may search for pages in your database that are similar to any other web page whether it's in the database or not. The URL for finding similar pages has the form shown below.

```
http://www.mysite.com/taxis/search/~>
  ↪similar.html?pr=default&ref=http://somesite/somepage.html
```

If the profile to be searched is "default" the `pr=default&` portion may be omitted:

```
ref=http://somesite/somepage.html
```

If the profile to be searched is anything other than "default" that must be specified instead of `default`:

```
pr=myprofile&ref=http://somesite/somepage.html
```

If the page to be located is the page the URL is on the `ref=URL` portion may be omitted:

```
/taxis/search/similar.html
or
/taxis/search/similar.html?pr=myprofile
```

The similar function will lookup the desired URL in the database or, if it's not in the database, fetch it from the webserver. It will then search the database looking for indexed pages similar to the specified page.

You could place a URL like this on all of your pages so users could, with one click, find all pages on your site similar in content to the one they were reading.

4.3 Using the Thesaurus Feature

You can create a thesaurus to either replace or add to the default thesaurus. The creation procedure is the same for either usage. Note that a thesaurus is not limited to synonyms. It can contain anything you wish to associate with a particular word: i.e., identities, generalities, or specifics of the word entry, plus associated phrases, acronyms, or spelling variations. The appliance maintains a collection of thesauri that you upload. For each profile you may select which, if any, thesaurus to use.

Here are the steps to use the thesaurus feature.

- Create a thesaurus file. Use the syntax described in the document “User Equivalence File Format” at the following url: http://www.thunderstone.com/site/texisman/~>user_equivalence_file_format.html

That document refers to the thesaurus as an “equivalence file”.

- Upload your thesaurus to the appliance. At the main menu click **Maintenance** then under **Search Appliance Settings** click **Custom Thesaurus**. The Custom Thesaurus page opens.
- In the **Name** field, enter a symbolic name that will be listed as an option in search settings. This name does not have to be related to the filename on disk in any way.
- In the **Permutations** field, choose a value. This value controls how many variations of your defined terms to create during indexing of your uploaded source file. Here is an example of the effect of the various values.

Assume a thesaurus entry of: car,ford,chevy,toyota

Permutation None: Just the terms as you entered them. Query “car” would find “car”, “ford”, “chevy”, and “toyota”. Query “ford” would only find “ford”.

Permutations Single: The terms you entered and the reverse. Same as above plus a query for any of “ford”, “chevy”, or “toyota” would find “car”.

Permutations Full: Equate every term with every other in each entry. Same as above plus a query for “ford” would find “chevy” and “toyota”.

- In the **New File** field, enter (or browse to) the file on your disk to upload. Click **Save Changes** to upload and index the file. When indexing is completed, you will receive a report about the indexing. If **Show results of indexing** is checked, you will also get a summary of the indexed words.
- After your thesaurus is installed on the appliance you can go to **Search Settings** for a profile to activate the thesaurus. There are three related options: **Synonyms**, **Main thesaurus**, and **Secondary Thesaurus**.
- Set **Synonyms** using the following information. **Synonyms** indicates how you want to apply a thesaurus (either yours or the default) to queries.

Disabled: no phrase recognition and no synonyms (equivalences)

Phrase recognition only: recognize query word groups that are known phrases and search for them as phrases

Phrases & Allow synonyms: phrase recognition plus allowing the tilde () operator to match synonyms on specific query terms

Phrases & Use synonyms by default: phrase recognition and matching synonyms on all query terms (tilde to turn off on specific terms).

- Set the Main Thesaurus and Secondary Thesaurus fields by using the following information. If you want to use only your thesaurus and not the default one, select yours for the Main Thesaurus option and leave verb 'Secondary Thesaurus' set to none. If you want the default in addition to your own, leave Main Thesaurus set to Built-In and set Secondary Thesaurus to yours. The names listed in these options are the symbolic names (Name field) you gave your thesauri when uploading them.
- Click Update to apply these settings. There is no need to check Apply Appearance, and these settings are applied to both Test Search and Live Search.

4.4 Getting Software Updates

You can obtain software updates manually or automatically. For information about getting them automatically, refer to Setup/Edit Update Preferences (3.11.2).

Use the following procedure to manually obtain software updates from Thunderstone. You are able to select which updates you want, if any.

- At main menu, click Maintenance.
- At Maintenance pane, click Check for updates.
- At proxy entry screen, enter proxy information if it is needed to reach the Internet, and then click Continue.
- A list of available updates is presented. Check the boxes for updates that you want to download from Thunderstone, and click Yes. The updates are downloaded to your Appliance, but they are not installed yet.
- A window for installation opens. Click Yes to install the updates.
- When the installation is completed, a message indicates that updates are completed.

4.5 Page Exclusion, Robots.txt, and Meta-robots

On the first access to a site the file /robots.txt will be retrieved, if its exists. Settings there will be respected. Any encountered URL that is disallowed by robots.txt will be discarded. Meta robots is also respected for each page retrieved. See <http://www.robotstxt.org/wc/exclusion.html> for the robots.txt and meta robots standards.

If there are any HTML trees that you don't want indexed you may want to setup a robots.txt file, meta robots within the HTML pages, or use the various exclusion options to the Search Appliance. For example:

if you had a “text only” version of your web server that duplicated the content of your normal server you would not want to index it. (On the other hand if most of your meaningful text is contained in graphics, Java, or JavaScript you may want to walk the text tree instead of the normal one, since graphics and Java are not searchable.)

Suppose your “text only” pages were all under a directory called `/text`. The simplest way to prevent traversal of that tree would be to use the exclusion or exclusion prefix.

The exclusion would look something like this:

```
/text/
```

The exclusion prefix would look something like this:

```
http://www.mysite.com/text/
```

That will prevent retrieval of any pages under the `/text` tree. This does not prevent other Web robots from retrieving the `/text` tree. To setup a permanent global exclusion list you need to create a file called `robots.txt` in your document root directory. The format of that file is as follows:

```
User-agent: *
Disallow: /text
```

Where `*` is the name of the robot to block. `*` means any robot not specifically named (all robots in this case since no others are named). Or you could specify the name of the robot. For the Appliance it would be `ThunderstoneSA`. You may specify several “Disallow”s for any given robot (see below). The “Disallow”s are simple path prefixes. They may not contain wildcards.

You may also specify different “Disallow” sets for different robots. Simply insert a blank line and add another “User-agent” line followed by its “Disallow” lines.

Here’s a larger example:

```
User-agent: *
Disallow: /text
Disallow: /junk

User-agent: ThunderstoneSA
Disallow: /text
Disallow: /thunderstonesa

User-agent: Scooter
Disallow: /text
Disallow: /junk
Disallow: /big
```

The `Scooter` robot will be blocked from accessing any pages under the `/text`, `/junk`, and `/big` trees. The Search Appliance will be blocked from accessing any pages under `/text` and `/thunderstonesa`. All other robots will be blocked from accessing pages under `/text` and `/junk`.

Use of `robots.txt` is not enforced in any way. Robots may or may not use it. The Search Appliance will, by default, always look for it and use it if present. This may be disabled by turning off “Respect robots.txt”. When using `robots.txt` you may still use “Exclusions” for manual exclusion.

Meta robots provides another method of controlling robots such as the Search Appliance. Any HTML may contain a meta tag in the source of the form.

```
<meta name="robots" content="WHAT-TO-DO">
```

WHAT-TO-DO may contain any of the following keywords. Multiple keywords may be used by placing a comma(,) between them.

Table 4.1: Meta-Robots Flags

Keyword	Meaning
INDEX	Index the text of this page
NOINDEX	Don't index the text of this page
FOLLOW	Follow hyperlinks on this page
NOFOLLOW	Don't follow hyperlinks on this page
ALL	Synonym for INDEX,FOLLOW
NONE	Synonym for NOINDEX,NOFOLLOW

Like `robots.txt` this is not enforced in any way. Robots may or may not use it. The Search Appliance always indexes and follows hyperlinks by default so it only looks for NOINDEX and/or NOFOLLOW and/or NONE.

4.6 Indexing Other Sites

You may index a site other than your own by specifying its URL just as you would for your own site.

```
http://www.anothersite.com
```

Please be kind when indexing other sites. Many are low bandwidth or heavily used already and won't appreciate being hit hard.

4.7 Indexing Individual Pages

To add an individual HTML page to the database, but not go after any of its references, add it to the Single Page list box.

4.8 Reindexing on a Schedule

It is often desirable to reindex a given site on a regular basis because of continuously changing content. You may specify a `Rewalk Schedule` to handle this for you.

It is also useful to perform a single rewalk at a later time or date to avoid overloading a web server during heavy use periods.

4.9 Checking for Web Server Errors

When you start a walk you will be sent to the walk status page. You may also reach that page at any time by selecting `Walk Status` from the menu. This page will show you the summary status of the running walk. When the walk completes you will see a summary of the walk as well as a list of any errors encountered. Following the error list is a list of duplicate pages encountered.

You may also view document linkage and info and errors from the `List/Edit URLs` page (3.3.6) from the menu.

4.10 Removing Pages from the Database

Use the `List/Edit URLs` menu (3.3.6) to find and delete specific URLs from the the database. You may delete individual pages or many pages at once using wildcards.

4.11 Erasing the Entire Database

If you decide to wipe out your existing database and it's settings to start over go to "Profiles" and click "Delete" next to the profile you wish to delete. This will completely remove the selected walk database and all options related to it.

4.12 Using Multiple Databases

Once you have a live searchable database you may want to build a separate one to contain different kinds of pages or to experiment with, without destroying your live database. Use the `Profiles` menu to create a new profile and database. You create the new profile with default settings or with a copy of the settings from another profile.

4.13 Integrating the Search Appliance with your Site

There are three main techniques to integrate the Search Appliance with your site. The techniques are categorized as follows:

- Static Host
- Dynamic Host and HTML
- Dynamic Host and XML

The simplest technique, *Static Host*, uses the built-in capability of the appliance to present a search page directly to a site visitor. Although this technique can be used with a dynamic host, it is commonly used with a static host. On your site, you present either a search field or a simple link. If you present a search field, when a visitor submits a query, the query is sent to the appliance. If you present a link, when a visitor clicks the link, a search page generated by the appliance is presented to the visitor, and the visitor uses this search page to submit a query. In either case, after a query is sent to the appliance, the appliance responds by sending the search results (HTML) to the visitor's browser. Note that you can customize the HTML of the search page, and this allows you to maintain a consistent appearance for your site and the search page generated by the appliance.

The *Dynamic Host and HTML* technique can be used in dynamically generated web sites. The host server sends a search query (http request) to the appliance, and the appliance responds by sending search results as HTML to the host server. The host server is responsible for sending the search query to the appliance, handling the HTML search results from the appliance, and for all interactions with the site visitor.

The *Dynamic Host and XML* technique can be used in dynamically generated web sites. The host server sends a search query (http request) to the appliance, and the appliance responds with the search results as XML. The host server is responsible for sending the search query to the appliance, handling the XML search results from the appliance, and for all interactions with the site visitor.

4.13.1 Static Host

Use the information in this section to perform the *Static Host* type of integration.

- Decide whether your existing pages will include a query field or just a link to the appliance search page.
- If your page will use a query field (an HTML form), you can obtain the HTML code that you need from the appliance's live search page as follows:

On the appliance, in the Administrative Interface, at the main menu, click *Live Search*. This opens the search form.

Use your web browser's view page source option (MSIE: TopMenu->View->Source, Netscape or Mozilla: TopMenu->View->Page Source) to open a window that contains the source code of the page.

Cut everything between and including the `<FORM>` and `</FORM>`. Paste the form into your web page(s).
- If your page will just link to the appliance search page, create the link using the url of the Test Search. To obtain the url of the Test Search, at the Administrator Interface, on the main menu, click *Test Search*. When the appliance search page opens, cut the url string from your browser, and paste it into your web page(s) at the appropriate link element.

- Optionally, if you want to change the appearance of the search page, you can do this by adding HTML to the Top and Bottom HTML settings. At the Administrator Interface main menu, click the Search Settings link, and scroll until Top HTML and Bottom HTML are in view. Add the desired HTML code. For information about using an HTML editor to make these code additions, refer to (section 3.6.19).
- The settings are ready for test runs. Use Update button but do not check Apply Appearance checkbox yet.
- After you are satisfied with the appearance and operation of searches using Test Search, you are ready to go live.
- On the Search Settings page, check the Apply Appearance checkbox and click Update button.
- If you are using a link to the appliance search page, change the link to point to the Live Search, using the same steps you used to set up the link to point to the Test Search.

4.13.2 Dynamic Host and HTML

Issuing a Query Programmatically

Use the information in this section to issue a query programmatically.

You can use either POST or GET to issue the search query. The only required variables are *pr* (profile), *query* and *dropXSL*. Variables not specified in the query take default values.

Here is an example URL for a search.

```
http://SEARCHAPPLIANCE/taxis/search/main.xml?dropXSL=0&pr=default~&
  &prox=page&rorder=500&rprox=500&rdfreq=500&rwfreq=500~&
  &rlead=500&sufs=2&order=r&query=query
```

Where *SEARCHAPPLIANCE* is the IP/hostname of your search appliance, and *query* is the user's query.

The following table provides a description of the query variables.

Processing Search Results

The appliance returns search results as HTML, so the results can be passed along to the site visitor without changes, or they can be modified or expanded before they are sent.

4.13.3 Dynamic Host and XML

This section provides information about issuing a query programatically and receiving the XML search results from the appliance.

Table 4.2: Search Query Variables

Variable	Description
pr	Specifies the search appliance profile.
prox	Proximity: words should be in the same line, sentence, paragraph, or page.
rorder	Word order: terms in same order as query are better (0-1000; 500 = medium).
rprox	Indicates how close to each other the words need to be (0-1000; 500 = medium).
rdfreq	Controls importance of frequency in the table (0-1000; 500 = medium).
rwfreq	Controls importance of frequency in the document (0-1000; 500 = medium).
rlead	Controls importance of closeness to document start (0-1000; 500 = medium).
sufs	Controls word forms (suffix). Values are 0 (exact), 1 (plurals), or 2 (any).
order	Controls the sort order. Values are relevance (r) or date (d).
query	Search query entered by site visitor.
cq	(Not shown in example) For categories. cq=1 for 1st, cq=2 for 2nd etc.
tq	(Not shown in example) Used for title-only queries.
uq	(Not shown in example) Used for URL Prefix queries.
dq	(Not shown in example) Used for depth queries.
sq	(Not shown in example) Used for site-specific queries.
sr	(Not shown in example) Used to limit results per site.

Issuing a Query Programmatically

You can use either POST or GET to issue the search query. The only required variables are `pr` (profile), `query` and `dropXSL`. Variables not specified in the query take default values.

Here is an example URL for a search.

```
http://SEARCHAPPLIANCE/taxis/search/main.xml?dropXSL=1&pr=default~
  ↳&prox=page&rorder=500&rprox=500&rdfreq=500&rwfreq=500~
  ↳&rlead=500&sufs=2&order=r&query=query
```

Where *SEARCHAPPLIANCE* is the IP/host of your search appliance, and *query* is the user's query.

Refer to [Issuing a Query Programmatically 4.13.2](#) for definitions of the query variables.

Processing Search Results

Settings in the administrative interface `Search Settings` control the format of the data the appliance returns in response to a query. The `Allow XML` value must be `Yes` to obtain search results as XML. Set the `Results Style` to `XSL Style` so `dropXSL` does not use the XSLT, and the appliance will just return the raw XML.

The XML elements are described in `XML Elements in Search Results` (section 5.7).

Sample ASP Code

The following ASP code demonstrates sending an http GET command to the appliance and receiving XML search results from the appliance.

```
<%
  on error resume next

  dim objSrvHTTP
  dim objXMLSend
  dim objXMLReceive

  set objSrvHTTP = Server.CreateObject("MSXML2.ServerXMLHTTP.4.0")
  set objXMLSend = Server.CreateObject("MSXML2.DOMDocument.4.0")
  set objXMLReceive = Server.CreateObject("MSXML2.DOMDocument.4.0")

  if err.number <> 0 then
    Response.Write err.description
    Response.Write "First error in code."
  end if
  err.clear

  objXMLSend.async = false
  objXMLSend.loadXML("<msg><id>2</id></msg>")
  objSrvHTTP.open
  "GET", "http://SearchAppliance/teaxis/search/main.xml?dropXSL=1&
pr=test&prox=page&rorder=500&rprox=500&rdfreq=500&rwfreq=500&
rlead=500&sufs=0&query=test&submit=Submit", false

  if err.number <> 0 then
    Response.Write err.description
    Response.Write "Second error in code."
  end if
  err.clear

  objSrvHTTP.send objXMLSend
  set objXMLReceive = objSrvHTTP.responseXML
  Response.ContentType = "text/xml"
  Response.Write objXMLReceive.xml

  if err.number <> 0 then
    Response.Write err.description
    Response.Write "Third error in code."
  end if
  err.clear
%>
```

4.14 Using Best Bets

The Search Appliance allows you to create links that will appear either at the top or to the right of the search results when specific keywords are searched for. They can be used for suggested links, or to promote specific URLs so they stand out from the main results. The Best Bet links are arranged into groups, which allow you to enable or disable a group of results easily.

The first step in create best bet links is to define a group. This is done from the “Group Settings” tab. You can name the group, and decide which information will be displayed about the group.

After creating a group you can add keywords to specific URLs. From the “List/Edit URLs” page enter the URL you want, and click on the URL to get the details on that URL. Currently you can only use URLs that have been walked and are in the database. There is a form on the page that allows you to add keywords to that URL. You can define a priority, title, description, group and keywords for the URL.

If the users query matches the keywords then the Best Bet will be shown. If several Best Bets match the query the highest priority is shown first. The title, description and URL are shown according to the group setting. The title and description can contain HTML code. Be careful that it does not disrupt the rest of the page layout. You can create multiple entries for the same URL. Each time you save a new set of blank boxes will be shown.

Once the Best Bets are created you can go to the “Search Settings” page to set up how they are displayed. For the top and right placements you can define which group is shown there, what title if any to display above the links, and the color, size and style of the boxes around the Best Bets.

As with any of the Search Settings these will apply to the “Test Search” first, and then when you apply the settings be copied to the “Live Search”, allowing you to test the settings and make sure they are appropriate before going live.

4.15 Using Access Control

The concepts and actions of access control in the Search Appliance are discussed in detail in the Access Control section, p. 65. The following are some general tips on how to setup and maintain access control rights.

4.15.1 Initial Lockdown

Since the default mode for Access Control when created is to allow all rights to all users for back-compatibility, it is recommended that perms be “locked down” first, and only granted as needed. The admin user, having the irrevocable ability to reset ACLs, should remain a “superuser” with all access, and other accounts turned into lesser-permission users. Lockdown should happen in this order:

1. Allow superuser: The admin user should have an Allow entry for all rights to the top-level Global object¹.

¹In version 5.3.0 and earlier, the admin user should instead be explicitly granted all rights to each of the second-level objects (All Users, All Groups, All Profiles, All Settings, and Maintenance).

2. Deny everyone: The group Everyone should have a Deny entry for all rights to the top-level Global object.

With these perms, users other than admin – including new users and profiles created in the future – will not be able to see or modify administrative settings. They can be granted perms as needed later, for example, the Read right could be removed from the Global deny ACE so that they can read but not modify any admin action/setting.

4.15.2 Example: User with Complete Control on One Profile

To configure a user that has complete access to just one specific profile (but no other profiles, nor the rest of administration such as creating accounts etc.), set up the lockdown settings above, then:

1. Create a Profile ACE on the specific profile, for that user, read and write access, and type Allow.
2. Create a Setting ACE for All Settings, for that user, read, write and delete access, type Allow.

The user will now be able to modify any setting on that profile, as well as start/stop walks on it, but will not be able to edit other profiles.

4.15.3 Example: User with Look and Feel Control on All Profiles

To configure a user that has the ability to change the Top and Bottom HTML on *any* profile, but cannot edit walk settings, nor start nor stop a walk, etc., set up the lockdown settings above, then:

1. Create a Profile ACE on All Profiles, for that user, read and write access, and type Allow.
2. Create a Setting ACE for Top HTML, for that user, read, write and delete access, type Allow.
3. Create a Setting ACE for Bottom HTML, for that user, read, write and delete access, type Allow.

The user will now be able to change the top and bottom HTML for any profile.

4.16 Indexing File Servers

The Thunderstone Search Appliance can index Windows and Unix file servers in addition to web servers. To do so you will first need to mount the file server to the appliance, and then configure the walk. To mount the file server you need to go to the Maintenance page and select Network Filesystems & Shares3.11.4.

After you've mounted your server to the appliance the interface will give you the base url to use in your profile. You may use that url or anything underneath it.

4.17 Replication

4.17.1 Replication Overview

In replication, a server profile sends walk data to another server profile. The two profiles can be on different machines or they can be on the same one. If the profiles are on different machines, the sending and receiving profiles can have the same or different names. If the profiles are on the same machine, use different profile names.

Here is an example that illustrates the replication process. In this example, the `Sender` profile has been set up as the sender profile and `Receiver` is the receiver profile. After `Sender` performs a walk, it sends the walk data to `Receiver`. The `Receiver` profile accepts the data as-is, without regard to its own profile settings. Only the profile that performed the walk may send the walk data, so in this example `Receiver` cannot replicate (the data it received from `Sender`) to another profile.

To avoid undesired overwriting of replication walk data, you should not allow the receiver profile to perform walks.

Before the receiver will accept replication data, the sender(s) need to be granted permission to send the data. This permission is managed in a cluster member list.

A good use of replication is to set up multiple machines to replicate to a single receiving profile. For example, machines A, B, and C each have a different profile, and they each replicate their walk data to a profile on machine D, which is the receiver. Another use of replication is to send walk data from multiple profiles on a machine to a single receiver profile that is on the same machine. This provides a means of combining walk data into a single profile. Another use of replication is to replicate data from one sender to multiple receivers. This way multiple machines hold the same walk data.

4.17.2 Procedure

The procedure in this section is an example of setting up replication on a single machine. It can be adapted to multiple machine configurations by changing the Replication Settings.

Set up the Sender Profile

- Choose an existing, walkable profile to be the sender. Or go to the `Profiles` menu item and create one, filling in all fields for a normal walk. We'll assume this profile is called `Sender`.
- Go to the `All Walk Settings` menu item for the `Sender` profile.
- Scroll down to `Replication Settings`.
- Enter the information for the receiver. In this example, `Host IP or Name` is `localhost` because we'll be sending data to the same machine, and `Profile Name` is `Receiver`. The page now includes the location of the receiver profile.
- Click `Update and Go` button.

- After a moment, the `Walk Status` page opens. Notice that there are N items in the replication queue. The number N is similar to the number of pages that were walked. The items remain in the queue, because they cannot be sent until the receiver profile is created (below). Normally, when a receiver profile is present, the contents of the queue are automatically sent to the receiver.

Create the Receiver Profile

- Create a new profile called `Receiver` via the `Profiles` menu item. (This matches the receiving profile name we entered on the `Sender` profile.)
- At main menu click `Maintenance`, then under `Search Appliance Settings` heading, click `System Wide Settings`.
- At the `Cluster Members` field, enter the IP address for each server that will send walk data to this machine. Use a separate line for each entry. In this example, there is one sending IP address, and it is `127.0.0.1` (use IP numbers, not the word `localhost`). To enable an entire subnet to send data, use an IP prefix and wildcard, eg. `10.10.*`.
- Click `Update` button.
- At main menu, click `Profiles`.
- When `Profiles` page opens, click `Sender`. A `Walk Settings` page opens for the `Sender` profile.
- Click `Walk Status` button. The `Walk Status` page for the `Sender` profile opens.
- There are still N items in the replication queue.
- Click the `replication queue` link.
- The items in the replication queue are sent to the `Receiver` profile. On the `Walk Status` page, there are now 0 items in the replication queue, which indicates the items were sent.
- On main menu, click `Profiles`, click `Receiver`, click `Walk Status` and observe that there is a list of pages recently walked. These pages were not walked by `Receiver`, instead they were obtained from `Sender`, which performed the walk.

4.17.3 DataLoad API

The replication system can also be used to load data directly onto the Search Appliance from an outside source, instead of “pulling” it from a URL or its links. This can be used for data that is not permanently stored at its URL (eg. generated data), and therefore cannot be fetched for indexing; it can instead be pushed to the Search Appliance for indexing. This feature requires version 5.4.19 or later of the `taxisScripts` package (see `Maintenance / Check for updates`).

Before loading data onto the Search Appliance, it must be configured to accept data from the IP address(es) that will be sending to it. This procedure is the same as for replication; see the `Cluster Members` setting, p. 97.

Submission Format

Data is submitted to the Search Appliance with an HTTP POST request sent to a similar URL as the admin interface (eg. `http://.../dowalk`), but with `/recvdata.xml` appended. Eg.:

`http://www.mysite.com/taxis/dowalk/recvdata.xml`

The following POST variables must be set in the request. Be sure to URL-encode the values:

- `profile`
Set to the name of the receiving profile.
- `data`
Set to an XML document containing the data, and what to do with it (insert/delete/etc.). See below for details.

The data XML document has the following format. Be sure to HTML-encode values:

```
<ThunderstoneReplication>
  <Item>
    <Type>I</Type>
    <Size>150369</Size>
    <Visited>2005-10-25 15:25:18</Visited>
    <Dlsecs>0</Dlsecs>
    <Depth>0</Depth>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Title>Sprocket Specifications</Title>
    <Body>...</Body>
    <Keywords>sprockets, gears, hubs</Keywords>
    <Description>Sprocket details</Description>
    <Meta></Meta>
    <Category>Mechanical</Category>
    <Modified>2005-10-25 11:21:07</Modified>
    <NextCheck>2005-10-25 16:25:18</NextCheck>
    <Views>0</Views>
    <Clicks>0</Clicks>
    <CTR>0.000000</CTR>
    <Pop>0</Pop>
    <Charset>UTF-8</Charset>
    <Refs dt:dt="bin.base64">...</Refs>
    <Errors dt:dt="bin.base64">...</Errors>
  </Item>
</ThunderstoneReplication>
```

Any element whose text data might not be XML-safe (eg. binary chars in the `<Body>`) should be base64-encoded, and the attribute `dt:dt="bin.base64"` set in the tag. Eg. the `<Refs>` and `<Errors>` elements' text data are always base64-encoded.

The elements are:

- `<Type>` The action to take with this data. Text value may be one of:
 - `I` Insert the data (overwrite previous data for URL if any)
 - `D` Delete the URL
 - `DP` Delete the URL as a pattern (eg. `http://www.mysite.com/dir/*`)
 - `UI` Update search indexes (call after a batch of inserts/deletes)
- `<Size>` The integer size of the original document.
- `<Visited>` When the document was fetched, in `YYYY-MM-DD HH:MM:SS` format.
- `<Dlsecs>` Number of seconds taken to download the document.
- `<Depth>` Depth of URL from a Base URL, eg. 0 is a Base URL, 1 is one click away, etc.
- `<Url>` The URL of the document.
- `<Title>` The title of the document.
- `<Body>` The formatted body of the document.
- `<Keywords>` Any keywords for the document.
- `<Description>` The description of the document.
- `<Meta>` Any meta data for the document.
- `<Category>` The category the document is in, if any. Must be a category name from the profile's Categories.
- `<Modified>` The Last-Modified date of the document, in `YYYY-MM-DD HH:MM:SS` format.
- `<NextCheck>` When the document should be refreshed, in `YYYY-MM-DD HH:MM:SS` format.
- `<Views>` Number of views of the document: how many times it's been shown in search results.
- `<Clicks>` Number of clicks of the document: how many times it's been clicked on in search results.
- `<CTR>` Click-through-ratio: floating-point number ratio of clicks to views.
- `<Pop>` Document popularity: number of references (links) to it.
- `<Charset>` Character set of `<Body>` data. Should correspond with `Storage Charset` profile setting (p. 39). If a charset other than the `Storage Charset` is used, it should be a standard IANA charset that the Search Appliance can convert to the `Storage Charset`.
- `<Refs>` Optional element with references (child links) of the document.
- `<Errors>` Optional element with errors of the document.

The optional `<Refs>` element lists the links (references) from the given document, for parent-child linking. Its text value is a base64-encoded XML document with the following format when decoded:

```
<results xmlns:dt="urn:schemas-microsoft-com:datatypes">
  <result>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Ref>http://www.mysite.com/dir/otherpage.html</Ref>
  </result>
  ...
</results>
```

Each `<Url>` should be the same as the `<Url>` in the above `<Item>` block. The `<Ref>` is a single link from the page. Only one `<Ref>` may be listed per `<result>`; additional links should be sent with additional `<result>` elements.

The optional `<Errors>` element contains any errors to be logged for the document. Note that this may be empty or not present if no errors are to be logged. Its text value is a base64-encoded XML document with the following format when decoded:

```
<results xmlns:dt="urn:schemas-microsoft-com:datatypes">
  <result>
    <Url>http://www.mysite.com/dir/page.html</Url>
    <Reason>Document not found: 404 (Not Found)</Reason>
  </result>
  ...
</results>
```

As with the `<Refs>` element, the `<Url>` must correspond with the original `<Item>` `<Url>`, and multiple errors must be listed in separate `<result>` elements.

Reply Format

The response to a DataLoad request is an XML document:

```
<ThunderstoneReplicationResult>
  <ItemResult>
    <rid>000000000</rid>
    <Type>I</Type>
    <DP>1</DP>
    <Status>OK</Status>
  </ItemResult>
  <Rows>1</Rows>
  <Version>Version 5.01.1234567890 20051010 (...)  
    2005-10-10 12:34:56</Version>
</ThunderstoneReplicationResult>
```

The elements are:

- <rid> The replication id. Ignored.
- <Type> The action type specified in the request.
- <DP> The number of URLs deleted by a <Type>DP</Type> action. Element is not present for other <Type>.
- <Status> Result code:
 - OK Success
 - FAIL_UNKNOWNTYPE The <Type> was not recognized
 - NODATA No parseable data in request
 - Not Allowed Sender is not a Cluster Member
 - No Profile No profile set in request POST
 - FAIL Failed, unknown reason
- <Rows> How many request <Item>s were processed.
- <Version> Version and release date of the software.

Once data has been successfully loaded onto the Search Appliance, if the profile has any receiver profiles defined under Replication Settings, the data will also be queued for replication to those receivers.

4.18 Additional Fields

4.18.1 Overview

The additional fields feature in the the Search Appliance allows you to define structured data that can be searched on, sorted by, and included in the results when using an XSL stylesheet. Typical uses might include having prices, dates or ratings associated with the documents.

4.18.2 Populating

To populate the additional fields they should first be defined in the additional fields section of the settings. You can specify a name, which is used as the name of the XML element when displaying the results, as well as when using the DataLoad API.

Once the field has been defined it can be populated either via the DataLoad API or through the Data From Field section. The fields are positionally numbered, and you can load Extra Field 1, 2 and/or 3 from the page that is read. If you are loading from a META field you will typically want a search of `. +` and the Meta Name you are loading from.

4.18.3 Sorting

To sort the results you can use the `order` form variable. To specify the first field you can set the value to `af1`, for the second `af2` and for the third `af3`. If you want to reverse the sort order you add a `d` to the value, i.e. `af1d`, `af2d`, `af3d`.

4.18.4 Searching

To add a search restriction to the query you can specify form variables with a name constructed as `af#OP`, where `#` is the number of the additional field, 1, 2, or 3, and `OP` is one of the following:

- `eq` - the field is equal to the form variable (e.g. `af1eq`)
- `gt` - the field is greater than the form variable (e.g. `af2gt`)
- `gte` - the field is greater or equal to than the form variable
- `lt` - the field is less than the form variable
- `lte` - the field is less or equal to than the form variable

4.19 DBWalker

4.19.1 Overview

The Taxis DBWalker module provides a walkable HTML interface to a remote database. If there is a database server somewhere which has a JDBC driver, DBWalker can serve up that database via HTML, which can then be walked by the Search Appliance or viewed by users. DBWalker can be configured to print all records on a single page, or to provide an “index” page which creates links to individual pages, each of which shows a single record of the table.

DBWalker is different from the normal idea of an “import” in multiple ways:

- DBWalker does not do any actual “importing” at all - it simply enables a way to view parts of a database through a website. It’s still up to the Search Appliance to walk the given website and index its content.
- The idea of ‘import’ implies a one-time action. Because DBWalker provides a HTML interface, it can be used to keep up to date with changes to the remote database. If a single record in the remote database changes, then DBWalker’s HTML interface will be different, and a refresh crawl by the Search Appliance will see this and change its internal index accordingly.

JDBC connections are cached across HTTP requests. The first time a request for a configuration is received, it establishes a JDBC connection and keeps it for 5 minutes. If another request for the same configuration is received, it will re-use the same connection. This greatly enhances the Search Appliance’s crawling speed, and keeps from bogging down the remote database with unnecessary connect/disconnect activity. JDBC connections are closed after 5 minutes of inactivity.

4.19.2 Configuration Overview

DBWalker uses multiple individual configurations for the different databases and tables it needs to talk to. Each configuration describes a single group of settings for a single table in a single database. It is possible to have multiple configurations use the same table and databases – for example, you can have one configuration list the entire contents of a table, while another configuration limits the data to a certain range.

Each configuration specifies which database to talk to with a type (“PostgreSQL”, “Oracle”, etc., which determines what JDBC driver to use), a JDBC connection string (which specifies things like host, port, and database), a username, and a password. The configuration must also specify which table is to be read, and can optionally specify which columns to read (defaults to all), any filter for the data (by way of a “WHERE” clause), and a key field.

If no key field is specified, then DBWalker won’t know how to uniquely identify rows, so it will print all the data on a single HTML page. If a key field is specified, then DBWalker will create an index page, which lists only the key field column. Each row’s key field is listed as a link back to DBWalker, which will give a page displaying all of the selected fields of only that record. This allows more fine-grained indexing and searching in large tables.

4.19.3 DBWalker Output Overview

The DBWalker's internal libraries produce XML output. This is transformed into HTML via a XSL stylesheet, which is changeable on a per-configuration basis. When a request is received that ends in `.xml`, DBWalker will return an XML document with a reference to an XSL stylesheet on the DBWalker server (which modern browsers will automatically fetch and apply).

However, if a request is received with the extension `.html`, DBWalker will apply the stylesheet server-side before the client ever sees it, and hand the resulting HTML to the client. This is useful for clients that do not apply XSL transformations to XML documents (like the Search Appliance).

4.19.4 DBWalker Authentication Overview

There are two ways to do authentication with the remote database. Authentication information can be stored in the config file, or it can be provided dynamically.

If a username and password are provided in the configuration, then that user/pass will be used for every request for that config. This has the advantage that users never have to input a username/password, but also has the security disadvantage that anyone who opens the website can see the data. Depending on the contents of the database, this may or may not be significant.

The other option is to not include a username or password in the configuration. When DBWalker is invoked by in this situation, it will prompt for a username/password (via Basic authentication). If the remote database accepts the credentials, the page is displayed. For the Search Appliance to walk these pages, it will have to know a valid username/password for the pages. This is supplied in the `Login Info` section of `All Walk Settings` ((42)).

This integrates well with Results Authorization. if the Search Appliance's search is set to use Authorized Search Results with "Prompt via Form", the credentials will automatically be verified with DBWalker.

To summarize a few key points:

- If you include a username and password in the DBWalker config, anyone will be able to see the results, including searches with `Results Authorization` in use: any user's login will work since the correct user/pass is "built in" to the DBWalker config.
- If no user/pass is included in the config, then users will have to supply their own username/password, and the config can be used properly with `Results Authorization`.
- If no user/pass is included in the config, be sure to put a valid username/password in the `Login Info` (pg. 42) section of any profile that uses it so the Search Appliance is able to index the content. If the Search Appliance can't see the results, then no searches will find it!

4.19.5 Obtaining DBWalker

DBWalker is obtained through the `Check for Updates` section of appliance maintenance. The `j2re` update must be installed prior to the DBWalker update. Please see the `Getting Software Updates` section (pg. 86) for details on installing software updates.

4.19.6 Managing DBWalker

- At main menu, click Maintenance.
- At Maintenance pane, click DBWalker Settings / Status.

The main DBWalker Administration interface is divided into 3 sections:

- DBWalker Status

This shows whether the DBWalker is currently enabled, and if so what the process ID is. A button is provided to enable or disable the DBWalker, whichever is applicable.

A link is also provided to the DBWalker Global Options page.

- DBWalker Configurations

This section lists all the configurations available for DBWalker. It shows the configuration name, what type of database it uses, its table, and its JDBC connection string. If the DBWalker Server is currently running, the name of each configuration will be a link to that configuration's DBWalker page.

Here you have the option to edit or delete any configuration, and to create new ones with the "Create New Configuration..." button at the bottom. Please see the Managing DBWalker Configurations section below for more information on managing configurations.

- DBWalker Stylesheets

This section allows you to manage the XSL stylesheets that DBWalker uses. You can add, delete, or modify, or view stylesheets here. Please see the Managing DBWalker Stylesheets section (pg. 109) below for more information on stylesheets and how to manage them.

4.19.7 DBWalker Global Options

The DBWalker Global Options page lists settings that affect all configurations, usually involving the DBWalker environment.

These are considered "advanced" settings and should only need to be changed if advised to by Thunderstone Support.

- LogLevel

This affects how much information is written to DBWalker.log. Each level includes all levels above it.

- SEVERE - Errors that cause the DBWalker to fail, giving it no chance to continue. These include errors reading the server config file, server socket errors, errors setting up the JDBC classloader, etc.
- WARNING - errors that cause an individual request to fail, but allow the server to continue on servicing other requests. These include individual connection socket errors, individual configuration errors errors, unexpected SQL errors, etc.

- INFO - This is the default logging level. Logs errors that are probably caused by malformed clients, or other things that we think an admin should know about. Includes clients giving malformed HTTP headers, requesting nonexistent configs, server startup/shutdown notification, etc.
- CONFIG - reports all information being read from configuration files.
- FINE - More fine-tuned book-keeping without going into details about individual classes/methods. Includes worker/socket assignment, worker pool manipulation, cache manipulation, enter/exit of JDBC methods.
- FINER - reports run/stop/resume of individual threads, more details of worker processing of socket.
- FINEST - Kitchen sink and then some. This will cause the log to fill very quickly with superfluous information during normal operation, and is advised not to be used unless requested by Thunderstone Support.

The default value of INFO is fine for normal operation.

- Max JVM Memory

This allows you to increase the maximum amount of memory, in megabytes, that the JVM will allow itself to allocate. If you are working with very large tables and getting `OutOfMemoryException` errors, then you may need to increase this. The default value is 64Mb.

Note that this value is not the amount of memory that will be immediately allocated by the JVM - it will only allocate as much as it needs. This simply provides an upper limit on how much memory will be used.

4.19.8 Managing DBWalker Configurations

Choosing to either edit a configuration or create a new one takes you to a listing page where you can change the facets of a configuration.

- General Information

The General Information section contains things that don't pertain directly to the remote database itself.

- Configuration Name
If you're creating a new configuration, you will be asked to enter a name. It is used when specifying which group of settings you want to use when DBWalker is invoked, but has no bearing beyond that. Names may contain letters, numbers, dashes, and underscores (no spaces).
- Stylesheet
Specifies which XSL stylesheet to use. You can only use stylesheets that you've already uploaded. Please see the Managing DBWalker Stylesheets section (pg. 109) for more information.

- Max Rows per Page

Sets a maximum number of rows to use on a single index page. If there are more rows than is allowed on a single page, next and back links are used as necessary to see the rest of the links. This is because if a table contains 10 million rows, just generating the index page can take huge amounts of time. DBWalker can be told to only deal with 100 rows at a time, keeping it from getting bogged down.

- Appliance Link

If you are using an internal interface to access the Search Appliance's administrator interface, this can allow you to force the DBWalker to be walked through an interface that will be visible to external users. Usually the default for this will be fine.

If administrators are accessing the search appliance through an internal-only interface, let's say `internalonly.mysite.com`, then the DBWalker will get walked as `http://internalonly.mysite.com/taxis...` This will work fine for the walk itself, but when external users use search, they will see results referencing `internalonly.mysite.com`, which they won't be able to access.

By setting Appliance Link to something like `www.mysite.com/taxis` (or whatever external users will be able to see), then the DBWalker will get walked with the proper links.

- Database Information

The database information section collects information about how to connect to your remote database.

- Type

This determines which JDBC driver will be loaded. DBWalker comes with support for Oracle, Microsoft SQL Server, Sybase, PostgreSQL, and Taxis.

The Oracle (dedicated) type is used to connect to an Oracle database through dedicated mode instead of the default shared mode. There is a slight performance disadvantage to this, and should only be used when the ordinary Oracle type does not work.

Alternatively, you can select [jdbcConnect] as the type, which lets you manually enter the JDBC Connection String. The Host, Port, and DB/Service values are all contained in the JDBC connection string, so the Connect String field replaces all 3 of them.

- Host / Connect String The contents of this field depend on what Type you have selected.

Database Type	field contents
Oracle, Sybase, PostgreSQL, or MS SQL Server	the hostname of the machine you're connecting to, or its IP address.
Taxis	the hostname and full path to the jdbc script on the remote server, i.e. <code>host.mysite.com/taxis/jdbc</code> .
[jdbcConnect]	the full JDBC connection string.

If the type is [jdbcConnect], then this field is Connect String, which lets you specify the full JDBC connection string. This is useful if you already know the JDBC connection string for your remote DB and don't want to have to break it down into hostname, port, etc. The exact formatting of this string differs for each remote database type.

- Port The port number that the remote database is listening to.

Oracle, Sybase, PostgreSQL, or MS SQL Server	the port to use, or leave blank for the default.
Taxis	unused, already specified as part of the Host field.
[jdbcConnect]	unused, already specified as part of the Connect String field.

- DB/Service The contents of this field is dependent on your database type.

Sybase, MS SQL Server, or PostgreSQL	the name of the database you want to connect to.
Oracle	the name of the service to connect to.
Taxis	the full path to the remote database, i.e. C:\morph3\taxis\testdb\ or /var/db/testdb.
[jdbcConnect]	unused, already specified as part of the Connect String field.

- Username

The username to give to the remote database. If this is left blank, username/password will be asked from the user when a request is made. Please see the “DBWalker Authentication Overview” section (pg. 104) for more information.

If connecting to a Microsoft SQL Server database, it’s possible to enter DOMAIN/user as the username to use domain authentication, where DOMAIN is the domain that the server belongs to.

- Password

The password to give to the remote database. If this is left blank, username/password will be asked from the user when a request is made. Please see the “DBWalker Authentication Overview” section (pg. 104) for more information.

- Table Information

The table information section collects information about the table that you want to access.

- Table

the name of the SQL table you want to retrieve data from.

- Fields

An optional list of fields to retrieve from the table. By default, all fields are retrieved. This is specified as a comma-seperated list, as you would use in the beginning of a SQL query.

- Where clause

Allows you to limit the data returned by DBWalker. It is not limited to using the fields specified in the fields section. The where clause should not contain the SQL keyword WHERE, just the conditional clause. For example, if your table has an id and a name, you could set Fields to name and Where clause to id>100 to only get names of records where the id is greater than 100.

- Key Field

Specifies the “key” field of the database. This field should be able to uniquely identify each record in the table, allowing DBWalker to create a list of links to each record from a single index page. If no key field is specified, the entire contents of the table will be displayed.

4.19.9 Managing DBWalker Stylesheets

XSL Stylesheets allow you to customize the way the DBWalker results are presented via HTML. They are applied either server-side or client-side, as detailed in the DBWalker Output Overview section (pg. 104) above.

Here you can edit or delete stylesheets, or upload new ones. If the DBWalker server is running, you can view any stylesheet by clicking that stylesheet's name.

- **Uploading XSL Stylesheets**

Beneath the list of current stylesheets is the Upload Stylesheet section. Here you can browse to a .xsl file and upload it to the appliance.

If the file already exists, the Overwrite existing box must be checked when you upload, to make sure you acknowledge the old file will be overwritten.

- **Editing XSL Stylesheets**

There are two methods for edit XSL stylesheets. The simplest way is to click on the “Edit” button next to a stylesheet. This provides a page with a large text area that contains the contents of the stylesheet. Make your changes and click “Save” to save the changes.

If you're doing heavy development to a stylesheet, you'll probably find working in a text area very limiting. Alternatively you can download the stylesheet (by clicking on its name on the main DBWalker page), make the changes you want locally with your preferred XSL editor, and re-upload the file. If further tweaks are necessary, simply change the file and re-upload it as much as necessary.

4.19.10 Adding Configurations to Profiles

To get the contents of a DBWalker config searchable, you add it to one or more profiles, where it will be crawled with the rest of the profile.

On the All Walk Settings page (pg. 32), there is a multi-select box listing all of DBWalker's configurations. Simply select all the configurations you would like to be included in that profile, and they will be walked.

4.20 Thunderstone Proxy Module

4.20.1 Overview

The Thunderstone Proxy Module allows you to use built-in Windows authentication while performing searches. It accepts NTLM authentication, which Internet Explorer can be configured to automatically pass along. The proxy then passes the request along to the Search Appliance, which communicates with the proxy module to authorize the results.

This allows for authenticated searches, without requiring the users to enter their credentials redundantly. Once the Proxy Module is installed and configured (on a remote box), have users use the Proxy Module

machine's web server for searching instead of going to the Search Appliance directly. An `Authenticated Search` link is created in the `Start Menu` shortcuts upon installation.

4.20.2 Requirements

The Thunderstone Proxy Module must be installed on a machine with IIS 6, which is only available on Windows 2003. This must be a separate machine from the Search Appliance, and, if authenticating against a domain, the machine the proxy module is installed on must be a member of that domain.

Ensure that the machine that the Thunderstone Proxy Module is being installed on is a secure machine. Because the Proxy Module is dealing with authorization functionality, a user with Administrative privileges could potentially tamper with operations.

The Proxy Module machine must have the port range 1701-1799 open for incoming connections – but only from the Search Appliance box, not search users' browsers – in addition to port 80 to allow IIS to serve searches for users on the machine.

4.20.3 Installing the Proxy Module

Before installing the proxy module the only thing you need to know is:

- The full hostname of the Search Appliance machine (eg. `thunderstone.mysite.com`) that the Proxy Module will be communicating with.

You can download the proxy module from the Search Appliance machine by going to the Maintenance section, selecting `Extra Downloads`, then `Thunderstone Proxy Module`, and finally click the `Download proxyModuleInstaller.exe` link for the installer. Obviously, once downloaded, the installer must be run on the Windows 2003 machine you wish to become the proxy, in order to be installed.

When installing you will be asked for a few items:

- `Destination Location` - This is where the actual DLL for the proxy module and its supporting files are placed. The directory `windows\system32\inet_srv` is recommended by default.
- `VirtualDir Path` - This is the path that will be used for the IIS virtual directory that the proxy module is assigned to. Its actual location does not matter, as the proxy module will intercept all requests, but IIS still requires that all virtual directories point to a real path. The directory `Program Files\Thunderstone Software\Thunderstone Proxy Module` is suggested by default.
- `Hostname` - The full hostname of the Search Appliance machine that this Proxy Module should connect to.

4.20.4 Configuring the Search Appliance

There are three things that must be done in the Search Appliance to configure it to accept authentication information from the Thunderstone Proxy Module, one of them global and one on a per-profile basis.

Add the Proxy Machine to Cluster Members

The IP address of the machine that the Thunderstone Proxy Module is installed on must be added to the list of `Cluster Members` to tell the Search Appliance to trust the proxy machine.

- Choose `Maintenance` on the left.
- Choose `System Wide Settings`, under the `Search Appliance Settings` section.
- Enter the proxy machine's IP address in the `Cluster Members` field on a new line.

Enable Results Authorization for the Target Profile

Results Authorization must be enabled for the target profile, if it's not already enabled.

- Select the profile in the `Profiles` page.
- Choose `Search Settings` on the left.
- Set the radio button for `Authorization Method` to `Basic/NTLM/file` (occurs beneath `Login Cookies` and `Login URL`).
- Click `Update` at the bottom.

Make the Target Profiles Visible

The profiles that you want to search with the proxy module must be set `Visible`, which enables the profile for things like metasearching and the proxy module.

- Select the profile in the `Profiles` page.
- Choose `Search Settings` on the left.
- Set the `Visible` setting to `Y`.
- Click `Update` at the bottom.

Configuring Browsers for Passing Credentials

Even with the Proxy Module and the Search Appliance set up for automatic authentication, modern browsers do not automatically provide your credentials, for security reasons. They need to be configured to trust the sites in order to unobtrusively pass credentials.

Configuring Internet Explorer

Internet Explorer will only send credentials to sites if the site is listed in the Local Internet zone. The following steps adds the Proxy Module machine to IE's Local Internet:

- Start Internet Explorer.
- Choose Tools from the menu, and select Internet Options.
- Choose the Security tab.
- Choose Local Internet from the list of zones.
- Click the Sites button to edit the local internet.
- Click Advanced to manually add a site.
- Uncheck Require server verification (https:) for all sites in this zone
- Enter the full hostname of your proxying machine, for example proxyMachine.example.com.
- Click Add to add the site to the Trusted Sites.
- Click Close to close the 'verb'Advanced' window.
- Click OK to close the 'verb'Local Intranet' window.
- Click OK to close the Internet Options window.

Internet Explorer is now configured to pass credentials to the proxy machine.

Configuring Firefox and Mozilla on Windows

Firefox and Mozilla can also be configured to automatically pass along credentials, but they too must be told to trust the proxy machine.

Note: These settings have been verified in Firefox 1.5 and Mozilla Seamonkey 1.0.7, versions prior to this may not be supported.

- Start Firefox/Mozilla.
- Enter about:config in the address bar.
- Find the network.automatic-ntlm-auth.trusted-uris item and double click on it.
- Enter your proxy machine's full hostname, for example proxyMachine.example.com.
 - The list can be comma-separated for multiple machines.
- Click OK to close the Enter String Value window.

Firefox/Mozilla is now configured to automatically pass credentials to the proxying machine.

4.20.5 Manually Configuring the Proxy Module

This section describes how to manually configure IIS for use of the Thunderstone Proxy Module. This is **not** necessary for normal operations - these actions are performed automatically by InstallShield upon installation. These steps are only necessary if IIS's configuration gets wiped out and needs to be redone.

Overview

The Thunderstone Proxy Module is an ISAPI Extension that is assigned as a Global Application Map to a Virtual Directory `/taxis` in IIS. All requests to the `/taxis` are not be served from the file system that the virtual directory points to, but instead go through the Proxy Module, which sends the request off to the Search Appliance machine.

The Virtual Directory that the Proxy Module uses needs to be set to not deny anonymous access and only integrated authentication.

4.20.6 Manual Installation Steps

These are the steps that must be done if you are manually setting up IIS for using the Proxy Module. **Note that these are done automatically by the InstallShield wizard** for you and do *not* need to be manually done under normal circumstances.

- Open the IIS Configuration
 - Right click on My Computer on the desktop.
 - Select Manage...
 - Open Services and Applications in the tree.
 - Open Internet Information Services.
 - Open Web Sites.
 - Select the website you want to add the Proxy Module to (most likely Default Web Site).
- Add a new virtual directory
 - Right click on the website and select New -> Virtual Directory...
 - The Virtual Directory Creation Wizard opens. Click Next>.
 - In the Alias box, enter `taxis` and click Next>.
 - In the Path box, enter the real physical path you want the virtual directory to map to, and click Next. the Proxy Module uses the directory `<INSTALLDIR>/etc/ISAPI-virtualdir` by default.

Note that it doesn't matter what directory is selected. This directory will never be used because all requests will be intercepted by the Proxy Module. The only reason a directory must be selected is because IIS insists that *all* virtual directories map to a real physical location.

- At the Virtual Directory Access Permissions screen, just click Next to complete the wizard, as we won't be using any of the permissions.
- Click Finish to complete the wizard and return to the Computer Management window.
- Apply the Proxy Module as a Wildcard Application Map
 - Right-click on the newly created virtual directory and select Properties.
 - The lower half of the properties window is labeled Application Settings. Click Create to make a custom set of application settings for this virtual directory.
 - After clicking Create, the Configuration should no longer be disabled. Click Configuration.
 - The lower half of the new Application Configuration window details Wildcard Application Maps, which is currently empty. Click Insert.
 - Next to the Executable field, click the Browse button and locate TaxisISAPI.dll, which is in the directory you installed the Proxy Module to.
 - * (The default location for this file in C:\windows\system32\inetsrv).
 - **Uncheck** the box next to Verify that file exists, and click OK.
 - TaxisISAPI.dll will now be in the list of Wildcard Application Maps. Click OK to close the Application Configuration window.
- Configure the virtual directory for authentication
 - While still in the taxis Properties window for the new virtual directory, Select the Directory Security tab.
 - In the top section, labelled Authentication and Access Control, click the Edit... button.
 - **Uncheck** Enable Anonymous Access and ensure that Integrated Windows Authentication is **checked**.
 - Click OK to close the Authentication Methods window.
 - Click OK to close the taxis Properties window.
- Add the Proxy Module to IIS' list of allowed extensions

By default IIS blocks all ISAPI extensions as a security measure. The Proxy Module must be explicitly allowed in IIS' configuration.

 - Back in the Computer Management window, open Web Service Extensions, underneath Internet Information Services.
 - The right side of the window should now have a list of rules. Right-click beneath the existing rules and select Add a new web service extension...
 - In the Extension Name field, enter Thunderstone Proxy Module.
 - Next to the Required files text area, click the Add... button.
 - Next to Path to file:, click Browse... and locate TaxisISAPI.dll, (just as in the previous set of instructions), and click OK to close the Add File dialog.

- Check the box next to Set extension status to Allowed, and click OK to close the window.

IIS is now set up properly to use the Proxy Module. Note that changes still need to be made to the Search Appliance, as detailed in the **Configuring the Search Appliance** section on page 111.

Chapter 5

Reference

5.1 Database and File Usage

The Search Appliance maintains a database that contains text from HTML pages, links to other pages, and a list of categories.

When the Search Appliance walker runs it creates a new database, under your specified data directory, to hold the new walk. It then dispatches a separate process for each web site it needs to visit and another to handle all of the “Single Pages”. Each of these retrieves all of the pages in it’s base list and stores the text of the HTML page to the `html` table and the hyperlinks to the `refs` table. All of the desirable URLs from the page that have not been seen before are placed into an internal “todo” list. After all of the base URLs are processed the process repeats with the internal todo list. When there’s nothing left in the todo list processing is complete.

Once all of the walking is complete the indices needed for searching are created on the data. Then the new database is flagged as the “live” one and the old database is deleted. Therefore your disk must have sufficient space for 2 complete databases plus temporary space used during the indexing step.

The databases are called `db1` and `db2`. The Search Appliance alternates between using these two names.

Note that the above applies to a walk type of `New`. During a walk type of `Refresh` only one database, the “live” one, is used.

The Search Appliance also maintains a file containing the detailed report for each walk. This file has the same name as the database with `.long` appended to the end. Also, a single file called `summary` is maintained with short summary information about the state of the database.

Given a data directory named `.../default` there may also be the following:

```
.../default/db1 an actual walk database
.../default/db2 an actual walk database
.../default/db1.long detailed walk report. Displayed when viewing Walk Status
.../default/db2.long detailed walk report. Displayed when viewing Walk Status
```

.../default/summary summary walk report. Displayed as Walk summary when viewing Walk Settings

Each setting has a record in the `options` table of the default database. See section 5.3 (p. 120) for the list of fields in the table. At each complete rewalk the current options settings are copied into an options table in the walk database. These options are not changed as settings are modified and are not otherwise used unless a search is performed setting the database with `db` instead of setting the profile with `pr`.

5.2 Walk Database Tables and Fields

Table 5.1: html table

Field	Description
id	Unique record id
Hash	Document hash for duplicate content detection
Size	Size of retrieved html document
Visited	The date the page was modified (or fetched if modified not set)
Dlsecs	The number of seconds to fetch the page
Depth	The number of URLs traversed to reach the page
Url	The URL of the real HTML page
Title	The Title of the page
Body	The textual content of the page, in UTF-8
Keywords	The <code>keywords</code> meta data from the page
Description	The <code>description</code> meta data from the page
Meta	Other meta data from the page, separated by newlines
Catno	List of categories to which the URL belongs
Modified	The date the page was modified
NextCheck	The date the page should next be refreshed
Views	The number of times this URL has been viewed (shown in results)
Clicks	The number of times this URL has been clicked (in results)
CTR	Click-through ratio
Pop	Popularity (number of pages linking to this page)
MimeType	MIME type of original page (future use)
Charset	Character set of original page (future use)

Table 5.2: refs table

Field	Description
Url	The URL of the HTML page
Ref	The URL of a reference (link) on the HTML page

Table 5.3: categories table

Field	Description
Catno	The number for the category
Url	The URL pattern for the category
Category	The name of the category

Table 5.4: error table

Field	Description
Url	The URL of the an HTML page that could not be retrieved
Reason	The reason it could not be retrieved
id	Unique record id (includes timestamp info).

Table 5.5: querylog table - (only used if query logging is enabled)

Field	Description
id	Contains the date and time of the query (unique record id)
Client	The hostname of the web client that performed the query
Query	The user's query as entered

5.3 Options Table Fields

These are the options table fields (maintained in the default database):

Table 5.6: options table

Field	Description
id	Unique id for the record
Profile	The name of the profile that the record belongs to
Name	The name of the setting
Type	The data type of the setting (always String)
String	The value of the setting
Int	Unused
Float	Unused
Strlist	Unused

5.4 Customizing the Search

You may make common changes to the Search Appliance's search appearance by using *Search Settings* from the administrative interface main menu.

5.5 Customizing the Walker

You may make many changes to the Search Appliance's walk behavior by using *Walk Settings* from the administrative interface main menu.

5.6 Third-Party Software

The Search Appliance may contain and utilize the following third-party software to enhance its functionality, depending on the version purchased. Note that your usage and rights to such third-party software may be governed by the appropriate licenses originating with that software, in addition to your License Agreement with Thunderstone - EPI for Thunderstone software.

5.6.1 Antiword

The *antiword* package is used by Thunderstone's *anytotx* plugin to handle Microsoft(R) Word files. It has been modified to work within *anytotx*'s installation and to extract meta information. Thunderstone's modified source may be obtained from

`ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the modified Antiword source. Sending a CD will require payment of shipping and handling charges by the requestor. *antiword* is governed by the terms of the GNU GPL, which is reproduced on p. 140.

5.6.2 Aspell

The GNU Project's *aspell* package is executed by (but not linked or compiled into) the Search Appliance for spell-checking and "Did you mean..." queries. Complete source code and documentation is available at `ftp://ftp.thunderstone.com/pub/epi-gpl/aspell-0.50.3.tar.gz` or `ftp://ftp.thunderstone.com/pub/epi-gpl/aspell-0.60.4.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the source. Sending a CD will require payment of shipping and handling charges by the requestor. *aspell* is governed by the terms of the GNU Lesser GPL, which is reproduced on p. 157.

5.6.3 Catdoc xls2csv

Catdoc's *xls2csv* program is used by Thunderstone's *anytotx* plugin to handle Microsoft(R) Excel(R) spreadsheet files. It has been modified to work within *anytotx*'s installation and to extract meta

information. Thunderstone's modified source may be obtained from `ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the modified Catdoc source. Sending a CD will require payment of shipping and handling charges by the requestor. Catdoc is governed by the terms of the GNU GPL, which is reproduced on p. 140.

5.6.4 Cole library

The `cole` library is used by Thunderstone's versions of `catdoc` and `antiword`. It has been modified to prevent extraneous printing. Thunderstone's modified source may be obtained from `ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the modified `cole` source. Sending a CD will require payment of shipping and handling charges by the requestor. The `cole` library is governed by the terms of the GNU GPL, which is reproduced on p. 140.

5.6.5 iconv

GNU `libiconv` may be used by Thunderstone's HTML processor to convert documents in certain character sets. GNU `libiconv` is not incorporated into Thunderstone's products but is a separate standalone program, called via `exec()` and writing/reading standard input/output. You may obtain complete source code and documentation for `libiconv` at `ftp://ftp.thunderstone.com/pub/epi-gpl/libiconv-1.9.2.tar.gz` or by contacting Thunderstone tech support and requesting a CD containing the GNU `libiconv` source. Sending a CD will require payment of shipping and handling charges by the requestor. GNU `libiconv` is governed by the terms of the GNU Library GPL, which is reproduced on p. 157.

5.6.6 JDBC drivers

Walking external databases with the DBWalker module may use one or more of the following drivers.

Oracle JDBC driver

Walking external Oracle databases with the DBWalker module may use the Oracle driver, subject to the license below:

ORACLE TECHNOLOGY NETWORK DEVELOPMENT AND DISTRIBUTION LICENSE AGREEMENT

"We," "us," and "our" refers to Oracle USA, Inc., for and on behalf of itself and its subsidiaries and affiliates under common control. "You" and "your" refers to the individual or entity that wishes to use the programs from Oracle. "Programs" refers to the software product you wish to download and use and program documentation. "License" refers to your right to use the programs under the terms of this agreement. This agreement is governed by the substantive and

procedural laws of California. You and Oracle agree to submit to the exclusive jurisdiction of, and venue in, the courts of San Francisco, San Mateo, or Santa Clara counties in California in any dispute arising out of or relating to this agreement.

We are willing to license the programs to you only upon the condition that you accept all of the terms contained in this agreement. Read the terms carefully and select the "Accept" button at the bottom of the page to confirm your acceptance. If you are not willing to be bound by these terms, select the "Do Not Accept" button and the registration process will not continue.

License Rights

We grant you a nonexclusive, nontransferable limited license to use the programs for purposes of developing your applications. You may also distribute the programs with your applications to your customers. If you want to use the programs for any purpose other than as expressly permitted under this agreement you must contact us, or an Oracle reseller, to obtain the appropriate license. We may audit your use of the programs. Program documentation is either shipped with the programs, or documentation may be accessed online at:
<http://otn.oracle.com/docs>

Ownership and Restrictions

We retain all ownership and intellectual property rights in the programs. You may make a sufficient number of copies of the programs for the licensed use and one copy of the programs for backup purposes.

You may not:

- use the programs for any purpose other than as provided above;
- distribute the programs unless accompanied with your applications;
- charge your end users for use of the programs;
- remove or modify any program markings or any notice of our proprietary rights;
- use the programs to provide third party training on the content and/or functionality of the programs, except for training your licensed users;
- assign this agreement or give the programs, program access or an interest in the programs to any individual or entity except as provided under this agreement;
- cause or permit reverse engineering (unless required by law for interoperability), disassembly or decompilation of the programs;
- disclose results of any program benchmark tests without our prior consent; or,
- use any Oracle name, trademark or logo.

Program Distribution

We grant you a nonexclusive, nontransferable right to copy and distribute the programs to your end users provided that you do not

charge your end users for use of the programs and provided your end users may only use the programs to run your applications for their business operations. Prior to distributing the programs you shall require your end users to execute an agreement binding them to terms consistent with those contained in this section and the sections of this agreement entitled "License Rights," "Ownership and Restrictions," "Export," "Disclaimer of Warranties and Exclusive Remedies," "No Technical Support," "End of Agreement," "Relationship Between the Parties," and "Open Source." You must also include a provision stating that your end users shall have no right to distribute the programs, and a provision specifying us as a third party beneficiary of the agreement. You are responsible for obtaining these agreements with your end users.

You agree to: (a) defend and indemnify us against all claims and damages caused by your distribution of the programs in breach of this agreements and/or failure to include the required contractual provisions in your end user agreement as stated above; (b) keep executed end user agreements and records of end user information including name, address, date of distribution and identity of programs distributed; (c) allow us to inspect your end user agreements and records upon request; and, (d) enforce the terms of your end user agreements so as to effect a timely cure of any end user breach, and to notify us of any breach of the terms.

Export

You agree that U.S. export control laws and other applicable export and import laws govern your use of the programs, including technical data; additional information can be found on Oracle's Global Trade Compliance web site located at:

<http://www.oracle.com/products/export/index.html?content.html>

You agree that neither the programs nor any direct product thereof will be exported, directly, or indirectly, in violation of these laws, or will be used for any purpose prohibited by these laws including, without limitation, nuclear, chemical, or biological weapons proliferation.

Disclaimer of Warranty and Exclusive Remedies

THE PROGRAMS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND. WE FURTHER DISCLAIM ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT.

IN NO EVENT SHALL WE BE LIABLE FOR ANY INDIRECT, INCIDENTAL, SPECIAL, PUNITIVE OR CONSEQUENTIAL DAMAGES, OR DAMAGES FOR LOSS OF PROFITS, REVENUE, DATA OR DATA USE, INCURRED BY YOU OR ANY THIRD PARTY, WHETHER IN AN ACTION IN CONTRACT OR TORT, EVEN IF WE HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. OUR ENTIRE LIABILITY FOR DAMAGES HEREUNDER SHALL IN NO EVENT EXCEED ONE THOUSAND DOLLARS (U.S. \ \$1,000).

No Technical Support

Our technical support organization will not provide technical support, phone support, or updates to you for the programs licensed under this agreement.

Restricted Rights

If you distribute a license to the United States government, the programs, including documentation, shall be considered commercial computer software and you will place a legend, in addition to applicable copyright notices, on the documentation, and on the media label, substantially similar to the following:

NOTICE OF RESTRICTED RIGHTS

"Programs delivered subject to the DOD FAR Supplement are 'commercial computer software' and use, duplication, and disclosure of the programs, including documentation, shall be subject to the licensing restrictions set forth in the applicable Oracle license agreement. Otherwise, programs delivered subject to the Federal Acquisition Regulations are 'restricted computer software' and use, duplication, and disclosure of the programs, including documentation, shall be subject to the restrictions in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987). Oracle USA, Inc., 500 Oracle Parkway, Redwood City, CA 94065."

End of Agreement

You may terminate this agreement by destroying all copies of the programs. We have the right to terminate your right to use the programs if you fail to comply with any of the terms of this agreement, in which case you shall destroy all copies of the programs.

Relationship Between the Parties

The relationship between you and us is that of licensee/licensor. Neither party will represent that it has any authority to assume or create any obligation, express or implied, on behalf of the other party, nor to represent the other party as agent, employee, franchisee, or in any other capacity. Nothing in this agreement shall be construed to limit either party's right to independently develop or distribute software that is functionally similar to the other party's products, so long as proprietary information of the other party is not included in such software.

Open Source

"Open Source" software - software available without charge for use, modification and distribution - is often licensed under terms that require the user to make the user's modifications to the Open Source software or any software that the user 'combines' with the Open Source

software freely available in source code form. If you use Open Source software in conjunction with the programs, you must ensure that your use does not: (i) create, or purport to create, obligations of us with respect to the Oracle programs; or (ii) grant, or purport to grant, to any third party any rights to or immunities under our intellectual property or proprietary rights in the Oracle programs. For example, you may not develop a software program using an Oracle program and an Open Source program where such use results in a program file(s) that contains code from both the Oracle program and the Open Source program (including without limitation libraries) if the Open Source program is licensed under a license that requires any "modifications" be made freely available. You also may not combine the Oracle program with programs licensed under the GNU General Public License ("GPL") in any manner that could cause, or could be interpreted or asserted to cause, the Oracle program or any modifications thereto to become subject to the terms of the GPL.

Entire Agreement

You agree that this agreement is the complete agreement for the programs and licenses, and this agreement supersedes all prior or contemporaneous agreements or representations. If any term of this agreement is found to be invalid or unenforceable, the remaining provisions will remain effective.

Last updated: 03/09/05

JTDS JDBC driver

Walking external databases with the DBWalker module may use the JTDS driver, for SQL Server(R) and Sybase(R) databases. This driver is governed by the GNU Lesser GPL; see p. 147.

PostgreSQL JDBC driver

Walking external databases with the DBWalker module may use the PostgreSQL driver. The license is reproduced below:

Copyright (c) 1997-2005, PostgreSQL Global Development Group
All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the

distribution.

3. Neither the name of the PostgreSQL Global Development Group nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

MySQL JDBC driver

Walking external databases with the DBWalker module may use the MySQL driver, governed by the GNU GPL; see p. 140.

5.6.7 ppt2html, msg2html

ppt2html and msg2html may be used by Thunderstone's anytotx document filter to convert Microsoft(R) PowerPoint and .msg files. Source is available at:

```
ftp://ftp.thunderstone.com/pub/epi-gpl/ppt2html.c
ftp://ftp.thunderstone.com/pub/epi-gpl/msg2html.c
ftp://ftp.thunderstone.com/pub/epi-gpl/msfilt.tar.gz
```

or by contacting Thunderstone tech support and requesting a CD containing the source. Sending a CD will require payment of shipping and handling charges by the requestor. ppt2html and msg2html are governed by the terms of the GNU GPL, which is reproduced on p. 140.

5.6.8 SSL/HTTPS plugin

This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (<http://www.openssl.org/>). Copyright ©1998-2002 The OpenSSL Project. All rights reserved. This product includes cryptographic software written by Eric Young (eay@cryptsoft.com). Copyright ©1995-1998 Eric Young. All rights reserved.

The OpenSSL toolkit stays under a dual license, i.e. both the conditions of the OpenSSL License and the original SSLeay license apply to the toolkit. See below for the actual license

texts. Actually both licenses are BSD-style Open Source licenses. In case of any license issues related to OpenSSL please contact openssl-core@openssl.org.

OpenSSL License

Copyright (c) 1998-2002 The OpenSSL Project. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. All advertising materials mentioning features or use of this software must display the following acknowledgment:
"This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit. (<http://www.openssl.org/>)"
4. The names "OpenSSL Toolkit" and "OpenSSL Project" must not be used to endorse or promote products derived from this software without prior written permission. For written permission, please contact openssl-core@openssl.org.
5. Products derived from this software may not be called "OpenSSL" nor may "OpenSSL" appear in their names without prior written permission of the OpenSSL Project.
6. Redistributions of any form whatsoever must retain the following acknowledgment:
"This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (<http://www.openssl.org/>)"

THIS SOFTWARE IS PROVIDED BY THE OpenSSL PROJECT ``AS IS'' AND ANY EXPRESSED OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE OpenSSL PROJECT OR ITS CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This product includes cryptographic software written by Eric Young (eay@cryptsoft.com). This product includes software written by Tim Hudson (tjh@cryptsoft.com).

Original SSLeay License

Copyright (C) 1995-1998 Eric Young (eay@cryptsoft.com)
All rights reserved.

This package is an SSL implementation written
by Eric Young (eay@cryptsoft.com).
The implementation was written so as to conform with Netscapes SSL.

This library is free for commercial and non-commercial use as long as the following conditions are aheared to. The following conditions apply to all code found in this distribution, be it the RC4, RSA, lhash, DES, etc., code; not just the SSL code. The SSL documentation included with this distribution is covered by the same copyright terms except that the holder is Tim Hudson (tjh@cryptsoft.com).

Copyright remains Eric Young's, and as such any Copyright notices in the code are not to be removed. If this package is used in a product, Eric Young should be given attribution as the author of the parts of the library used. This can be in the form of a textual message at program startup or in documentation (online or textual) provided with the package.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. All advertising materials mentioning features or use of this software must display the following acknowledgement:
"This product includes cryptographic software written by
Eric Young (eay@cryptsoft.com)"
The word 'cryptographic' can be left out if the rouines from the library being used are not cryptographic related :-).
4. If you include any Windows specific code (or a derivative thereof) from the apps directory (application code) you must include an acknowledgement: "This product includes software written by
Tim Hudson (tjh@cryptsoft.com)"

THIS SOFTWARE IS PROVIDED BY ERIC YOUNG ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED

WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

The licence and distribution terms for any publically available version or derivative of this code cannot be changed. i.e. this code cannot simply be copied and put under another distribution licence [including the GNU Public Licence.]

5.6.9 unrar

The Thunderstone file converter plugin (anytotx) may utilize Alexander L. Roshal's unrar utility to unpack RAR archive files (*.rar). The unrar utility is governed by the unRAR license reproduced below:

```
*****      *****      *****      unRAR - free utility for RAR archives
**   **   **   **   **   **   ~~~~~~
*****      *****      *****      License for use and distribution of
**   **   **   **   **   **   ~~~~~~
**   **   **   **   **   **   FREE portable version
                                   ~~~~~~
```

The source code of unRAR utility is freeware. This means:

1. All copyrights to RAR and the utility unRAR are exclusively owned by the author - Alexander Roshal.
2. The unRAR sources may be used in any software to handle RAR archives without limitations free of charge, but cannot be used to re-create the RAR compression algorithm, which is proprietary. Distribution of modified unRAR sources in separate form or as a part of other software is permitted, provided that it is clearly stated in the documentation and source comments that the code may not be used to develop a RAR (WinRAR) compatible archiver.
3. The unRAR utility may be freely distributed. No person or company may charge a fee for the distribution of unRAR without written permission from the copyright holder.
4. THE RAR ARCHIVER AND THE UNRAR UTILITY ARE DISTRIBUTED "AS IS". NO WARRANTY OF ANY KIND IS EXPRESSED OR IMPLIED. YOU USE AT YOUR OWN RISK. THE AUTHOR WILL NOT BE LIABLE FOR DATA LOSS, DAMAGES, LOSS OF PROFITS OR ANY OTHER KIND OF LOSS WHILE USING OR MISUSING THIS SOFTWARE.

5. Installing and using the unRAR utility signifies acceptance of these terms and conditions of the license.
6. If you don't agree with terms of the license you must remove unRAR files from your storage devices and cease to use the utility.

Thank you for your interest in RAR and unRAR.

Alexander L. Roshal

5.6.10 unzip

The Thunderstone file converter plugin (anytotx) may utilize Info-ZIP's unzip utility to unpack ZIP archive files (*.zip). The unzip software is governed by the Info-ZIP license reproduced below:

This is version 2002-Feb-16 of the Info-ZIP copyright and license.
The definitive version of this document should be available at
<ftp://ftp.info-zip.org/pub/infozip/license.html> indefinitely.

Copyright (c) 1990-2002 Info-ZIP. All rights reserved.

For the purposes of this copyright and license, "Info-ZIP" is defined as the following set of individuals:

Mark Adler, John Bush, Karl Davis, Harald Denker, Jean-Michel Dubois, Jean-loup Gailly, Hunter Goatley, Ian Gorman, Chris Herborth, Dirk Haase, Greg Hartwig, Robert Heath, Jonathan Hudson, Paul Kienitz, David Kirschbaum, Johnny Lee, Onno van der Linden, Igor Mandrichenko, Steve P. Miller, Sergio Monesi, Keith Owens, George Petrov, Greg Roelofs, Kai Uwe Rommel, Steve Salisbury, Dave Smith, Christian Spieler, Antoine Verheijen, Paul von Behren, Rich Wales, Mike White

This software is provided "as is," without warranty of any kind, express or implied. In no event shall Info-ZIP or its contributors be held liable for any direct, indirect, incidental, special or consequential damages arising out of the use of or inability to use this software.

Permission is granted to anyone to use this software for any purpose, including commercial applications, and to alter it and redistribute it freely, subject to the following restrictions:

1. Redistributions of source code must retain the above copyright notice, definition, disclaimer, and this list of conditions.
2. Redistributions in binary form (compiled executables) must reproduce the above copyright notice, definition, disclaimer, and this list of conditions in documentation and/or other materials provided with the distribution. The sole exception to this condition is

redistribution of a standard UnZipSFX binary as part of a self-extracting archive; that is permitted without inclusion of this license, as long as the normal UnZipSFX banner has not been removed from the binary or disabled.

3. Altered versions--including, but not limited to, ports to new operating systems, existing ports with new graphical interfaces, and dynamic, shared, or static library versions--must be plainly marked as such and must not be misrepresented as being the original source. Such altered versions also must not be misrepresented as being Info-ZIP releases--including, but not limited to, labeling of the altered versions with the names "Info-ZIP" (or any variation thereof, including, but not limited to, different capitalizations), "Pocket UnZip," "WiZ" or "MacZip" without the explicit permission of Info-ZIP. Such altered versions are further prohibited from misrepresentative use of the Zip-Bugs or Info-ZIP e-mail addresses or of the Info-ZIP URL(s).
4. Info-ZIP retains the right to use the names "Info-ZIP," "Zip," "UnZip," "UnZipSFX," "WiZ," "Pocket UnZip," "Pocket Zip," and "MacZip" for its own source and binary releases.

5.6.11 zlib

The Search Appliance utilizes the `zlib` compression library. Copyright ©1995-2003 Jean-loup Gailly and Mark Adler.

5.6.12 SpiderMonkey (JavaScript-C) Engine

The `libtxjs.*` library (Thunderstone JavaScript plugin) contains and utilizes the SpiderMonkey engine, as well as additional functionality.

The `txjs.tar` file contains context diffs (patches) to the Mozilla Project's SpiderMonkey (JavaScript-C) engine, version 1.5-rc4. Complete documentation and source code to the SpiderMonkey Engine is available at <http://www.mozilla.org/js/spidermonkey/>.

The patches in `txjs.tar` were created by Thunderstone Software LLC and apply to the core SpiderMonkey engine. They are provided for compliance with the Netscape Public License, which governs usage of the SpiderMonkey engine. A copy of the Netscape Public License is on p. 166. Note that the `libtxjs.*` library also contains other (Thunderstone) code.

5.6.13 PDF/anytotx plugin

Portions of this product Copyright 1996-2000 Glyph & Cog, LLC.

5.6.14 thttpd - throttling HTTP server

the Search Appliance's vhttpd web server is derived in part from thttpd, Copyright ©1995 by Jef Poskanzer jef@acme.com. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR AND CONTRIBUTORS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

5.6.15 RedHat Linux

The Search Appliance uses the RedHat Linux operating system, version 7.3, which is licensed under the GNU Public License, p. 140. See also

<http://www.redhat.com/licenses/thirdparty/eula.html> for more information.

5.6.16 Webmin

The Search Appliance uses the Webmin web-based system administration system for maintaining and configuring the operating system. Copyright ©Jamie Cameron All rights reserved. Complete source is available at: <http://www.webmin.com/>. The license is reproduced below:

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the developer nor the names of contributors may be used to endorse or promote products derived from this software

without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE DEVELOPER ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE DEVELOPER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

5.6.17 Java

The Search Appliance uses the Java 2 run-time environment developed by Sun Microsystems, Inc. to index third-party databases using JDBC drivers. This product includes code licensed from RSA Security, Inc. Some portions licensed from IBM are available at <http://oss.software.ibm.com/icu4j/> as well. The license agreement is reproduced below:

Sun Microsystems, Inc. Binary Code License Agreement

READ THE TERMS OF THIS AGREEMENT AND ANY PROVIDED SUPPLEMENTAL LICENSE TERMS (COLLECTIVELY "AGREEMENT") CAREFULLY BEFORE OPENING THE SOFTWARE MEDIA PACKAGE. BY OPENING THE SOFTWARE MEDIA PACKAGE, YOU AGREE TO THE TERMS OF THIS AGREEMENT. IF YOU ARE ACCESSING THE SOFTWARE ELECTRONICALLY, INDICATE YOUR ACCEPTANCE OF THESE TERMS BY SELECTING THE "ACCEPT" BUTTON AT THE END OF THIS AGREEMENT. IF YOU DO NOT AGREE TO ALL THESE TERMS, PROMPTLY RETURN THE UNUSED SOFTWARE TO YOUR PLACE OF PURCHASE FOR A REFUND OR, IF THE SOFTWARE IS ACCESSED ELECTRONICALLY, SELECT THE "DECLINE" BUTTON AT THE END OF THIS AGREEMENT.

1. LICENSE TO USE. Sun grants you a non-exclusive and non-transferable license for the internal use only of the accompanying software and documentation and any error corrections provided by Sun (collectively "Software"), by the number of users and the class of computer hardware for which the corresponding fee has been paid.

2. RESTRICTIONS. Software is confidential and copyrighted. Title to Software and all associated intellectual property rights is retained by Sun and/or its licensors. Except as specifically authorized in any Supplemental License Terms, you may not make copies of Software, other than a single copy of Software for archival purposes. Unless enforcement is prohibited by applicable law, you may not modify, decompile, or reverse engineer Software. Licensee

acknowledges that Licensed Software is not designed or intended for use in the design, construction, operation or maintenance of any nuclear facility. Sun Microsystems, Inc. disclaims any express or implied warranty of fitness for such uses. No right, title or interest in or to any trademark, service mark, logo or trade name of Sun or its licensors is granted under this Agreement.

3. LIMITED WARRANTY. Sun warrants to you that for a period of ninety (90) days from the date of purchase, as evidenced by a copy of the receipt, the media on which Software is furnished (if any) will be free of defects in materials and workmanship under normal use. Except for the foregoing, Software is provided "AS IS". Your exclusive remedy and Sun's entire liability under this limited warranty will be at Sun's option to replace Software media or refund the fee paid for Software.

4. DISCLAIMER OF WARRANTY. UNLESS SPECIFIED IN THIS AGREEMENT, ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT THESE DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

5. LIMITATION OF LIABILITY. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL SUN OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR SPECIAL, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF OR RELATED TO THE USE OF OR INABILITY TO USE SOFTWARE, EVEN IF SUN HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. In no event will Sun's liability to you, whether in contract, tort (including negligence), or otherwise, exceed the amount paid by you for Software under this Agreement. The foregoing limitations will apply even if the above stated warranty fails of its essential purpose.

6. Termination. This Agreement is effective until terminated. You may terminate this Agreement at any time by destroying all copies of Software. This Agreement will terminate immediately without notice from Sun if you fail to comply with any provision of this Agreement. Upon Termination, you must destroy all copies of Software.

7. Export Regulations. All Software and technical data delivered under this Agreement are subject to US export control laws and may be subject to export or import regulations in other countries. You agree to comply strictly with all such laws and regulations and acknowledge that you have the responsibility to obtain such licenses to

export, re-export, or import as may be required after delivery to you.

8. U.S. Government Restricted Rights. If Software is being acquired by or on behalf of the U.S. Government or by a U.S. Government prime contractor or subcontractor (at any tier), then the Government's rights in Software and accompanying documentation will be only as set forth in this Agreement; this is in accordance with 48 CFR 227.7201 through 227.7202-4 (for Department of Defense (DOD) acquisitions) and with 48 CFR 2.101 and 12.212 (for non-DOD acquisitions).

9. Governing Law. Any action related to this Agreement will be governed by California law and controlling U.S. federal law. No choice of law rules of any jurisdiction will apply.

10. Severability. If any provision of this Agreement is held to be unenforceable, this Agreement will remain in effect with the provision omitted, unless omission would frustrate the intent of the parties, in which case this Agreement will immediately terminate.

11. Integration. This Agreement is the entire agreement between you and Sun relating to its subject matter. It supersedes all prior or contemporaneous oral or written communications, proposals, representations and warranties and prevails over any conflicting or additional terms of any quote, order, acknowledgment, or other communication between the parties relating to its subject matter during the term of this Agreement. No modification of this Agreement will be binding, unless in writing and signed by an authorized representative of each party.

JAVATM 2 RUNTIME ENVIRONMENT (J2RE), STANDARD EDITION,
VERSION 1.4.1_X SUPPLEMENTAL LICENSE TERMS

These supplemental license terms ("Supplemental Terms") add to or modify the terms of the Binary Code License Agreement (collectively, the "Agreement"). Capitalized terms not defined in these Supplemental Terms shall have the same meanings ascribed to them in the Agreement. These Supplemental Terms shall supersede any inconsistent or conflicting terms in the Agreement, or in any license contained within the Software.

1. Software Internal Use and Development License Grant. Subject to the terms and conditions of this Agreement, including, but not limited to Section 4 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to

reproduce internally and use internally the binary form of the Software complete and unmodified for the sole purpose of designing, developing and testing your Java applets and applications intended to run on the Java platform ("Programs").

2. License to Distribute Software. Subject to the terms and conditions of this Agreement, including, but not limited to Section 4 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to reproduce and distribute the Software, provided that (i) you distribute the Software complete and unmodified (unless otherwise specified in the applicable README file) and only bundled as part of, and for the sole purpose of running, your Programs, (ii) the Programs add significant and primary functionality to the Software, (iii) you do not distribute additional software intended to replace any component(s) of the Software (unless otherwise specified in the applicable README file), (iv) you do not remove or alter any proprietary legends or notices contained in the Software, (v) you only distribute the Software subject to a license agreement that protects Sun's interests consistent with the terms contained in this Agreement, and (vi) you agree to defend and indemnify Sun and its licensors from and against any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of any and all Programs and/or Software. (vi) include the following statement as part of product documentation (whether hard copy or electronic), as a part of a copyright page or proprietary rights notice page, in an "About" box or in any other form reasonably designed to make the statement visible to users of the Software: "This product includes code licensed from RSA Security, Inc.", and (vii) include the statement, "Some portions licensed from IBM are available at <http://oss.software.ibm.com/icu4j/>".

3. License to Distribute Redistributables. Subject to the terms and conditions of this Agreement, including but not limited to Section 4 (Java Technology Restrictions) of these Supplemental Terms, Sun grants you a non-exclusive, non-transferable, limited license to reproduce and distribute those files specifically identified as redistributable in the Software "README" file ("Redistributables") provided that: (i) you distribute the Redistributables complete and unmodified (unless otherwise specified in the applicable README file), and only bundled as part of Programs, (ii) you do not distribute additional software intended to supersede any component(s) of the Redistributables (unless otherwise specified in the

applicable README file), (iii) you do not remove or alter any proprietary legends or notices contained in or on the Redistributables, (iv) you only distribute the Redistributables pursuant to a license agreement that protects Sun's interests consistent with the terms contained in the Agreement, (v) you agree to defend and indemnify Sun and its licensors from and against any damages, costs, liabilities, settlement amounts and/or expenses (including attorneys' fees) incurred in connection with any claim, lawsuit or action by any third party that arises or results from the use or distribution of any and all Programs and/or Software, (vi) include the following statement as part of product documentation (whether hard copy or electronic), as a part of a copyright page or proprietary rights notice page, in an "About" box or in any other form reasonably designed to make the statement visible to users of the Software: "This product includes code licensed from RSA Security, Inc.", and (vii) include the statement, "Some portions licensed from IBM are available at <http://oss.software.ibm.com/icu4j/>".

4. Java Technology Restrictions. You may not modify the Java Platform Interface ("JPI", identified as classes contained within the "java" package or any subpackages of the "java" package), by creating additional classes within the JPI or otherwise causing the addition to or modification of the classes in the JPI. In the event that you create an additional class and associated API(s) which (i) extends the functionality of the Java platform, and (ii) is exposed to third party software developers for the purpose of developing additional software which invokes such additional API, you must promptly publish broadly an accurate specification for such API for free use by all developers. You may not create, or authorize your licensees to create, additional classes, interfaces, or subpackages that are in any way identified as "java", "javax", "sun" or similar convention as specified by Sun in any naming convention designation.

5. Notice of Automatic Software Updates from Sun. You acknowledge that the Software may automatically download, install, and execute applets, applications, software extensions, and updated versions of the Software from Sun ("Software Updates"), which may require you to accept updated terms and conditions for installation. If additional terms and conditions are not presented on installation, the Software Updates will be considered part of the Software and subject to the terms and conditions of the Agreement.

6. Notice of Automatic Downloads. You acknowledge that, by your use of the Software and/or by requesting services that

require use of the Software, the Software may automatically download, install, and execute software applications from sources other than Sun ("Other Software"). Sun makes no representations of a relationship of any kind to licensors of Other Software. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL SUN OR ITS LICENSORS BE LIABLE FOR ANY LOST REVENUE, PROFIT OR DATA, OR FOR SPECIAL, INDIRECT, CONSEQUENTIAL, INCIDENTAL OR PUNITIVE DAMAGES, HOWEVER CAUSED REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF OR RELATED TO THE USE OF OR INABILITY TO USE OTHER SOFTWARE, EVEN IF SUN HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Trademarks and Logos. You acknowledge and agree as between you and Sun that Sun owns the SUN, SOLARIS, JAVA, JINI, FORTE, and iPLANET trademarks and all SUN, SOLARIS, JAVA, JINI, FORTE, and iPLANET-related trademarks, service marks, logos and other brand designations ("Sun Marks"), and you agree to comply with the Sun Trademark and Logo Usage Requirements currently located at: <http://www.sun.com/policies/trademarks>. Any use you make of the Sun Marks inures to Sun's benefit.

8. Source Code. Software may contain source code that is provided solely for reference purposes pursuant to the terms of this Agreement. Source code may not be redistributed unless expressly provided for in this Agreement.

9. Termination for Infringement. Either party may terminate this Agreement immediately should any Software become, or in either party's opinion be likely to become, the subject of a claim of infringement of any intellectual property right.

For inquiries please contact: Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, California 95054, U.S.A. (LFI#120080/Form ID#011801)

5.6.18 OpenSSL RPM

The Search Appliance uses a Perl module that contains OpenSSL. Copyright ©1996-2002 Sampo Kellomaki sampo@symlabs.com All Rights Reserved. See p. 127 for more information.

5.6.19 RAID utilities

The Search Appliance may use RAID utilities developed by the Adaptec Corporation. These are used by Thunderstone for system maintenance in this product. Usage is governed by the license below:

Copyright (c) 1996-2004, Adaptec Corporation

All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the Adaptec Corporation nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

5.6.20 GNU General Public License

Some third-party software packages shipped with the Search Appliance are governed by the GNU General Public License, reproduced below. See the Third-Party Software section, p. 121, for a list of applicable packages.

GNU GENERAL PUBLIC LICENSE
Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.
675 Mass Ave, Cambridge, MA 02139, USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users. This General Public License applies to most of the Free Software Foundation's software and to any other program whose authors commit to using it. (Some other Free Software Foundation software is covered by

the GNU Library General Public License instead.) You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid anyone to deny you these rights or to ask you to surrender the rights. These restrictions translate to certain responsibilities for you if you distribute copies of the software, or if you modify it.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must give the recipients all the rights that you have. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

We protect your rights with two steps: (1) copyright the software, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the software.

Also, for each author's protection and ours, we want to make certain that everyone understands that there is no warranty for this free software. If the software is modified by someone else and passed on, we want its recipients to know that what they have is not the original, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that redistributors of a free program will individually obtain patent licenses, in effect making the program proprietary. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

The precise terms and conditions for copying, distribution and modification follow.

GNU GENERAL PUBLIC LICENSE
TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License applies to any program or other work which contains a notice placed by the copyright holder saying it may be distributed under the terms of this General Public License. The "Program", below, refers to any such program or work, and a "work based on the Program" means either the Program or any derivative work under copyright law: that is to say, a work containing the Program or a portion of it, either verbatim or with modifications and/or translated into another language. (Hereinafter, translation is included without limitation in

the term "modification".) Each licensee is addressed as "you".

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running the Program is not restricted, and the output from the Program is covered only if its contents constitute a work based on the Program (independent of having been made by running the Program). Whether that is true depends on what the Program does.

1. You may copy and distribute verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and give any other recipients of the Program a copy of this License along with the Program.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Program or any portion of it, thus forming a work based on the Program, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) You must cause the modified files to carry prominent notices stating that you changed the files and the date of any change.
- b) You must cause any work that you distribute or publish, that in whole or in part contains or is derived from the Program or any part thereof, to be licensed as a whole at no charge to all third parties under the terms of this License.
- c) If the modified program normally reads commands interactively when run, you must cause it, when started running for such interactive use in the most ordinary way, to print or display an announcement including an appropriate copyright notice and a notice that there is no warranty (or else, saying that you provide a warranty) and that users may redistribute the program under these conditions, and telling the user how to view a copy of this License. (Exception: if the Program itself is interactive but does not normally print such an announcement, your work based on the Program is not required to print an announcement.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Program, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Program, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the

entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Program.

In addition, mere aggregation of another work not based on the Program with the Program (or with a work based on the Program) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may copy and distribute the Program (or a work based on it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you also do one of the following:

- a) Accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
- b) Accompany it with a written offer, valid for at least three years, to give any third party, for a charge no more than your cost of physically performing source distribution, a complete machine-readable copy of the corresponding source code, to be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange; or,
- c) Accompany it with the information you received as to the offer to distribute corresponding source code. (This alternative is allowed only for noncommercial distribution and only if you received the program in object code or executable form with such an offer, in accord with Subsection b above.)

The source code for a work means the preferred form of the work for making modifications to it. For an executable work, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the executable. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

If distribution of executable or object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place counts as distribution of the source code, even though third parties are not compelled to copy the source along with the object code.

4. You may not copy, modify, sublicense, or distribute the Program

except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense or distribute the Program is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

5. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Program or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Program (or any work based on the Program), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Program or works based on it.

6. Each time you redistribute the Program (or any work based on the Program), the recipient automatically receives a license from the original licensor to copy, distribute or modify the Program subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

7. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Program at all. For example, if a patent license would not permit royalty-free redistribution of the Program by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Program.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system, which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

8. If the distribution and/or use of the Program is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Program under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

9. The Free Software Foundation may publish revised and/or new versions of the General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of this License, you may choose any version ever published by the Free Software Foundation.

10. If you wish to incorporate parts of the Program into other free programs whose distribution conditions are different, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

11. BECAUSE THE PROGRAM IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

12. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A

FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

Appendix: How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the program's name and a brief idea of what it does.>
Copyright (C) 19yy <name of author>
```

```
This program is free software; you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation; either version 2 of the License, or
(at your option) any later version.
```

```
This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
GNU General Public License for more details.
```

```
You should have received a copy of the GNU General Public License
along with this program; if not, write to the Free Software
Foundation, Inc., 675 Mass Ave, Cambridge, MA 02139, USA.
```

Also add information on how to contact you by electronic and paper mail.

If the program is interactive, make it output a short notice like this when it starts in an interactive mode:

```
Gnomovision version 69, Copyright (C) 19yy name of author
Gnomovision comes with ABSOLUTELY NO WARRANTY; for details type
'show w'.
This is free software, and you are welcome to redistribute it
under certain conditions; type 'show c' for details.
```

The hypothetical commands 'show w' and 'show c' should show the appropriate parts of the General Public License. Of course, the commands you use may be called something other than 'show w' and 'show c'; they could even be mouse-clicks or menu items--whatever suits your program.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the program, if necessary. Here is a sample; alter the names:

Yoyodyne, Inc., hereby disclaims all copyright interest in the program 'Gnomovision' (which makes passes at compilers) written by James Hacker.

<signature of Ty Coon>, 1 April 1989
Ty Coon, President of Vice

This General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Library General Public License instead of this License.

5.6.21 GNU Lesser General Public License

Some third-party software packages distributed with the Search Appliance are governed by the GNU Lesser General Public License, reproduced below. See the Third-Party Software section, p. 121, for a list of applicable packages.

GNU LESSER GENERAL PUBLIC LICENSE
Version 2.1, February 1999

Copyright (C) 1991, 1999 Free Software Foundation, Inc.
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

[This is the first released version of the Lesser GPL. It also counts as the successor of the GNU Library Public License, version 2, hence the version number 2.1.]

Preamble

The licenses for most software are designed to take away your freedom to share and change it. By contrast, the GNU General Public Licenses are intended to guarantee your freedom to share and change free software--to make sure the software is free for all its users.

This license, the Lesser General Public License, applies to some specially designated software packages--typically libraries--of the Free Software Foundation and other authors who decide to use it. You can use it too, but we suggest you first think carefully about whether this license or the ordinary General Public License is the better strategy to use in any particular case, based on the explanations below.

When we speak of free software, we are referring to freedom of use,

not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for this service if you wish); that you receive source code or can get it if you want it; that you can change the software and use pieces of it in new free programs; and that you are informed that you can do these things.

To protect your rights, we need to make restrictions that forbid distributors to deny you these rights or to ask you to surrender these rights. These restrictions translate to certain responsibilities for you if you distribute copies of the library or if you modify it.

For example, if you distribute copies of the library, whether gratis or for a fee, you must give the recipients all the rights that we gave you. You must make sure that they, too, receive or can get the source code. If you link other code with the library, you must provide complete object files to the recipients, so that they can relink them with the library after making changes to the library and recompiling it. And you must show them these terms so they know their rights.

We protect your rights with a two-step method: (1) we copyright the library, and (2) we offer you this license, which gives you legal permission to copy, distribute and/or modify the library.

To protect each distributor, we want to make it very clear that there is no warranty for the free library. Also, if the library is modified by someone else and passed on, the recipients should know that what they have is not the original version, so that the original author's reputation will not be affected by problems that might be introduced by others.

Finally, software patents pose a constant threat to the existence of any free program. We wish to make sure that a company cannot effectively restrict the users of a free program by obtaining a restrictive license from a patent holder. Therefore, we insist that any patent license obtained for a version of the library must be consistent with the full freedom of use specified in this license.

Most GNU software, including some libraries, is covered by the ordinary GNU General Public License. This license, the GNU Lesser General Public License, applies to certain designated libraries, and is quite different from the ordinary General Public License. We use this license for certain libraries in order to permit linking those libraries into non-free programs.

When a program is linked with a library, whether statically or using a shared library, the combination of the two is legally speaking a combined work, a derivative of the original library. The ordinary General Public License therefore permits such linking only if the entire combination fits its criteria of freedom. The Lesser General Public License permits more lax criteria for linking other code with the library.

We call this license the "Lesser" General Public License because it does Less to protect the user's freedom than the ordinary General Public License. It also provides other free software developers Less of an advantage over competing non-free programs. These disadvantages are the reason we use the ordinary General Public License for many libraries. However, the Lesser license provides advantages in certain special circumstances.

For example, on rare occasions, there may be a special need to encourage the widest possible use of a certain library, so that it becomes a de-facto standard. To achieve this, non-free programs must be allowed to use the library. A more frequent case is that a free library does the same job as widely used non-free libraries. In this case, there is little to gain by limiting the free library to free software only, so we use the Lesser General Public License.

In other cases, permission to use a particular library in non-free programs enables a greater number of people to use a large body of free software. For example, permission to use the GNU C Library in non-free programs enables many more people to use the whole GNU operating system, as well as its variant, the GNU/Linux operating system.

Although the Lesser General Public License is Less protective of the users' freedom, it does ensure that the user of a program that is linked with the Library has the freedom and the wherewithal to run that program using a modified version of the Library.

The precise terms and conditions for copying, distribution and modification follow. Pay close attention to the difference between a "work based on the library" and a "work that uses the library". The former contains code derived from the library, whereas the latter must be combined with the library in order to run.

GNU LESSER GENERAL PUBLIC LICENSE

TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License Agreement applies to any software library or other program which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Lesser General Public License (also called "this License"). Each licensee is addressed as "you".

A "library" means a collection of software functions and/or data prepared so as to be conveniently linked with application programs (which use some of those functions and data) to form executables.

The "Library", below, refers to any such software library or work which has been distributed under these terms. A "work based on the Library" means either the Library or any derivative work under copyright law: that is to say, a work containing the Library or a

portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Source code" for a work means the preferred form of the work for making modifications to it. For a library, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the library.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Library is not restricted, and output from such a program is covered only if its contents constitute a work based on the Library (independent of the use of the Library in a tool for writing it). Whether that is true depends on what the Library does and what the program that uses the Library does.

1. You may copy and distribute verbatim copies of the Library's complete source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Library.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Library or any portion of it, thus forming a work based on the Library, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) The modified work must itself be a software library.
- b) You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.
- c) You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.
- d) If a facility in the modified Library refers to a function or a table of data to be supplied by an application program that uses the facility, other than as an argument passed when the facility is invoked, then you must make a good faith effort to ensure that, in the event an application does not supply such function or table, the facility still operates, and performs whatever part of its purpose remains meaningful.

(For example, a function in a library to compute square roots has

a purpose that is entirely well-defined independent of the application. Therefore, Subsection 2d requires that any application-supplied function or table used by this function must be optional: if the application does not supply it, the square root function must still compute square roots.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Library, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Library, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or collective works based on the Library.

In addition, mere aggregation of another work not based on the Library with the Library (or with a work based on the Library) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may opt to apply the terms of the ordinary GNU General Public License instead of this License to a given copy of the Library. To do this, you must alter all the notices that refer to this License, so that they refer to the ordinary GNU General Public License, version 2, instead of to this License. (If a newer version than version 2 of the ordinary GNU General Public License has appeared, then you can specify that version instead if you wish.) Do not make any other change in these notices.

Once this change is made in a given copy, it is irreversible for that copy, so the ordinary GNU General Public License applies to all subsequent copies and derivative works made from that copy.

This option is useful when you wish to copy part of the code of the Library into a program that is not a library.

4. You may copy and distribute the Library (or a portion or derivative of it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange.

If distribution of object code is made by offering access to copy from a designated place, then offering equivalent access to copy the

source code from the same place satisfies the requirement to distribute the source code, even though third parties are not compelled to copy the source along with the object code.

5. A program that contains no derivative of any portion of the Library, but is designed to work with the Library by being compiled or linked with it, is called a "work that uses the Library". Such a work, in isolation, is not a derivative work of the Library, and therefore falls outside the scope of this License.

However, linking a "work that uses the Library" with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a "work that uses the library". The executable is therefore covered by this License. Section 6 states terms for distribution of such executables.

When a "work that uses the Library" uses material from a header file that is part of the Library, the object code for the work may be a derivative work of the Library even though the source code is not. Whether this is true is especially significant if the work can be linked without the Library, or if the work is itself a library. The threshold for this to be true is not precisely defined by law.

If such an object file uses only numerical parameters, data structure layouts and accessors, and small macros and small inline functions (ten lines or less in length), then the use of the object file is unrestricted, regardless of whether it is legally a derivative work. (Executables containing this object code plus portions of the Library will still fall under Section 6.)

Otherwise, if the work is a derivative of the Library, you may distribute the object code for the work under the terms of Section 6. Any executables containing that work also fall under Section 6, whether or not they are linked directly with the Library itself.

6. As an exception to the Sections above, you may also combine or link a "work that uses the Library" with the Library to produce a work containing portions of the Library, and distribute that work under terms of your choice, provided that the terms permit modification of the work for the customer's own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the work that the Library is used in it and that the Library and its use are covered by this License. You must supply a copy of this License. If the work during execution displays copyright notices, you must include the copyright notice for the Library among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

- a) Accompany the work with the complete corresponding machine-readable source code for the Library including whatever

changes were used in the work (which must be distributed under Sections 1 and 2 above); and, if the work is an executable linked with the Library, with the complete machine-readable "work that uses the Library", as object code and/or source code, so that the user can modify the Library and then relink to produce a modified executable containing the modified Library. (It is understood that the user who changes the contents of definitions files in the Library will not necessarily be able to recompile the application to use the modified definitions.)

b) Use a suitable shared library mechanism for linking with the Library. A suitable mechanism is one that (1) uses at run time a copy of the library already present on the user's computer system, rather than copying library functions into the executable, and (2) will operate properly with a modified version of the library, if the user installs one, as long as the modified version is interface-compatible with the version that the work was made with.

c) Accompany the work with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 6a, above, for a charge no more than the cost of performing this distribution.

d) If distribution of the work is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.

e) Verify that the user has already received a copy of these materials or that you have already sent this user a copy.

For an executable, the required form of the "work that uses the Library" must include any data and utility programs needed for reproducing the executable from it. However, as a special exception, the materials to be distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of other proprietary libraries that do not normally accompany the operating system. Such a contradiction means you cannot use both them and the Library together in an executable that you distribute.

7. You may place library facilities that are a work based on the Library side-by-side in a single library together with other library facilities not covered by this License, and distribute such a combined library, provided that the separate distribution of the work based on the Library and of the other library facilities is otherwise permitted, and provided that you do these two things:

a) Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities. This must be distributed under the terms of the Sections above.

b) Give prominent notice with the combined library of the fact that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.

8. You may not copy, modify, sublicense, link with, or distribute the Library except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Library is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

9. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Library or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Library (or any work based on the Library), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Library or works based on it.

10. Each time you redistribute the Library (or any work based on the Library), the recipient automatically receives a license from the original licensor to copy, distribute, link with or modify the Library subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties with this License.

11. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Library at all. For example, if a patent license would not permit royalty-free redistribution of the Library by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Library.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply, and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any

patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

12. If the distribution and/or use of the Library is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Library under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

13. The Free Software Foundation may publish revised and/or new versions of the Lesser General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library specifies a version number of this License which applies to it and "any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Library does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

14. If you wish to incorporate parts of the Library into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

15. BECAUSE THE LIBRARY IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LIBRARY, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LIBRARY "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE

LIBRARY IS WITH YOU. SHOULD THE LIBRARY PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LIBRARY AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LIBRARY (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LIBRARY TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

How to Apply These Terms to Your New Libraries

If you develop a new library, and you want it to be of the greatest possible use to the public, we recommend making it free software that everyone can redistribute and change. You can do so by permitting redistribution under these terms (or, alternatively, under the terms of the ordinary General Public License).

To apply these terms, attach the following notices to the library. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the library's name and a brief idea of what it does.>
Copyright (C) <year> <name of author>
```

```
This library is free software; you can redistribute it and/or
modify it under the terms of the GNU Lesser General Public
License as published by the Free Software Foundation; either
version 2.1 of the License, or (at your option) any later version.
```

```
This library is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU
Lesser General Public License for more details.
```

```
You should have received a copy of the GNU Lesser General Public
License along with this library; if not, write to the Free Software
Foundation, Inc.,
59 Temple Place, Suite 330, Boston, MA 02111-1307 USA
```

Also add information on how to contact you by electronic and paper mail.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the library, if

necessary. Here is a sample; alter the names:

Yoyodyne, Inc., hereby disclaims all copyright interest in the library
'Frob' (a library for tweaking knobs) written by James Random Hacker.

<signature of Ty Coon>, 1 April 1990
Ty Coon, President of Vice

That's all there is to it!

5.6.22 GNU Library General Public License

Some third-party software packages distributed with the Search Appliance are governed by the GNU Library General Public License, reproduced below. See the Third-Party Software section, p. 121, for a list of applicable packages.

GNU LIBRARY GENERAL PUBLIC LICENSE
Version 2, June 1991

Copyright (C) 1991 Free Software Foundation, Inc.
59 Temple Place - Suite 330, Boston, MA 02111-1307, USA
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

[This is the first released version of the library GPL. It is
numbered 2 because it goes with version 2 of the ordinary GPL.]

Preamble

The licenses for most software are designed to take away your
freedom to share and change it. By contrast, the GNU General Public
Licenses are intended to guarantee your freedom to share and change
free software--to make sure the software is free for all its users.

This license, the Library General Public License, applies to some
specially designated Free Software Foundation software, and to any
other libraries whose authors decide to use it. You can use it for
your libraries, too.

When we speak of free software, we are referring to freedom, not
price. Our General Public Licenses are designed to make sure that you
have the freedom to distribute copies of free software (and charge for
this service if you wish), that you receive source code or can get it
if you want it, that you can change the software or use pieces of it
in new free programs; and that you know you can do these things.

To protect your rights, we need to make restrictions that forbid
anyone to deny you these rights or to ask you to surrender the rights.
These restrictions translate to certain responsibilities for you if
you distribute copies of the library, or if you modify it.

For example, if you distribute copies of the library, whether gratis or for a fee, you must give the recipients all the rights that we gave you. You must make sure that they, too, receive or can get the source code. If you link a program with the library, you must provide complete object files to the recipients so that they can relink them with the library, after making changes to the library and recompiling it. And you must show them these terms so they know their rights.

Our method of protecting your rights has two steps: (1) copyright the library, and (2) offer you this license which gives you legal permission to copy, distribute and/or modify the library.

Also, for each distributor's protection, we want to make certain that everyone understands that there is no warranty for this free library. If the library is modified by someone else and passed on, we want its recipients to know that what they have is not the original version, so that any problems introduced by others will not reflect on the original authors' reputations.

Finally, any free program is threatened constantly by software patents. We wish to avoid the danger that companies distributing free software will individually obtain patent licenses, thus in effect transforming the program into proprietary software. To prevent this, we have made it clear that any patent must be licensed for everyone's free use or not licensed at all.

Most GNU software, including some libraries, is covered by the ordinary GNU General Public License, which was designed for utility programs. This license, the GNU Library General Public License, applies to certain designated libraries. This license is quite different from the ordinary one; be sure to read it in full, and don't assume that anything in it is the same as in the ordinary license.

The reason we have a separate public license for some libraries is that they blur the distinction we usually make between modifying or adding to a program and simply using it. Linking a program with a library, without changing the library, is in some sense simply using the library, and is analogous to running a utility program or application program. However, in a textual and legal sense, the linked executable is a combined work, a derivative of the original library, and the ordinary General Public License treats it as such.

Because of this blurred distinction, using the ordinary General Public License for libraries did not effectively promote software sharing, because most developers did not use the libraries. We concluded that weaker conditions might promote sharing better.

However, unrestricted linking of non-free programs would deprive the users of those programs of all benefit from the free status of the libraries themselves. This Library General Public License is intended to permit developers of non-free programs to use free libraries, while

preserving your freedom as a user of such programs to change the free libraries that are incorporated in them. (We have not seen how to achieve this as regards changes in header files, but we have achieved it as regards changes in the actual functions of the Library.) The hope is that this will lead to faster development of free libraries.

The precise terms and conditions for copying, distribution and modification follow. Pay close attention to the difference between a "work based on the library" and a "work that uses the library". The former contains code derived from the library, while the latter only works together with the library.

Note that it is possible for a library to be covered by the ordinary General Public License rather than by this special one.

GNU LIBRARY GENERAL PUBLIC LICENSE

TERMS AND CONDITIONS FOR COPYING, DISTRIBUTION AND MODIFICATION

0. This License Agreement applies to any software library which contains a notice placed by the copyright holder or other authorized party saying it may be distributed under the terms of this Library General Public License (also called "this License"). Each licensee is addressed as "you".

A "library" means a collection of software functions and/or data prepared so as to be conveniently linked with application programs (which use some of those functions and data) to form executables.

The "Library", below, refers to any such software library or work which has been distributed under these terms. A "work based on the Library" means either the Library or any derivative work under copyright law: that is to say, a work containing the Library or a portion of it, either verbatim or with modifications and/or translated straightforwardly into another language. (Hereinafter, translation is included without limitation in the term "modification".)

"Source code" for a work means the preferred form of the work for making modifications to it. For a library, complete source code means all the source code for all modules it contains, plus any associated interface definition files, plus the scripts used to control compilation and installation of the library.

Activities other than copying, distribution and modification are not covered by this License; they are outside its scope. The act of running a program using the Library is not restricted, and output from such a program is covered only if its contents constitute a work based on the Library (independent of the use of the Library in a tool for writing it). Whether that is true depends on what the Library does and what the program that uses the Library does.

1. You may copy and distribute verbatim copies of the Library's complete source code as you receive it, in any medium, provided that

you conspicuously and appropriately publish on each copy an appropriate copyright notice and disclaimer of warranty; keep intact all the notices that refer to this License and to the absence of any warranty; and distribute a copy of this License along with the Library.

You may charge a fee for the physical act of transferring a copy, and you may at your option offer warranty protection in exchange for a fee.

2. You may modify your copy or copies of the Library or any portion of it, thus forming a work based on the Library, and copy and distribute such modifications or work under the terms of Section 1 above, provided that you also meet all of these conditions:

- a) The modified work must itself be a software library.
- b) You must cause the files modified to carry prominent notices stating that you changed the files and the date of any change.
- c) You must cause the whole of the work to be licensed at no charge to all third parties under the terms of this License.
- d) If a facility in the modified Library refers to a function or a table of data to be supplied by an application program that uses the facility, other than as an argument passed when the facility is invoked, then you must make a good faith effort to ensure that, in the event an application does not supply such function or table, the facility still operates, and performs whatever part of its purpose remains meaningful.

(For example, a function in a library to compute square roots has a purpose that is entirely well-defined independent of the application. Therefore, Subsection 2d requires that any application-supplied function or table used by this function must be optional: if the application does not supply it, the square root function must still compute square roots.)

These requirements apply to the modified work as a whole. If identifiable sections of that work are not derived from the Library, and can be reasonably considered independent and separate works in themselves, then this License, and its terms, do not apply to those sections when you distribute them as separate works. But when you distribute the same sections as part of a whole which is a work based on the Library, the distribution of the whole must be on the terms of this License, whose permissions for other licensees extend to the entire whole, and thus to each and every part regardless of who wrote it.

Thus, it is not the intent of this section to claim rights or contest your rights to work written entirely by you; rather, the intent is to exercise the right to control the distribution of derivative or

collective works based on the Library.

In addition, mere aggregation of another work not based on the Library with the Library (or with a work based on the Library) on a volume of a storage or distribution medium does not bring the other work under the scope of this License.

3. You may opt to apply the terms of the ordinary GNU General Public License instead of this License to a given copy of the Library. To do this, you must alter all the notices that refer to this License, so that they refer to the ordinary GNU General Public License, version 2, instead of to this License. (If a newer version than version 2 of the ordinary GNU General Public License has appeared, then you can specify that version instead if you wish.) Do not make any other change in these notices.

Once this change is made in a given copy, it is irreversible for that copy, so the ordinary GNU General Public License applies to all subsequent copies and derivative works made from that copy.

This option is useful when you wish to copy part of the code of the Library into a program that is not a library.

4. You may copy and distribute the Library (or a portion or derivative of it, under Section 2) in object code or executable form under the terms of Sections 1 and 2 above provided that you accompany it with the complete corresponding machine-readable source code, which must be distributed under the terms of Sections 1 and 2 above on a medium customarily used for software interchange.

If distribution of object code is made by offering access to copy from a designated place, then offering equivalent access to copy the source code from the same place satisfies the requirement to distribute the source code, even though third parties are not compelled to copy the source along with the object code.

5. A program that contains no derivative of any portion of the Library, but is designed to work with the Library by being compiled or linked with it, is called a "work that uses the Library". Such a work, in isolation, is not a derivative work of the Library, and therefore falls outside the scope of this License.

However, linking a "work that uses the Library" with the Library creates an executable that is a derivative of the Library (because it contains portions of the Library), rather than a "work that uses the library". The executable is therefore covered by this License. Section 6 states terms for distribution of such executables.

When a "work that uses the Library" uses material from a header file that is part of the Library, the object code for the work may be a derivative work of the Library even though the source code is not. Whether this is true is especially significant if the work can be

linked without the Library, or if the work is itself a library. The threshold for this to be true is not precisely defined by law.

If such an object file uses only numerical parameters, data structure layouts and accessors, and small macros and small inline functions (ten lines or less in length), then the use of the object file is unrestricted, regardless of whether it is legally a derivative work. (Executables containing this object code plus portions of the Library will still fall under Section 6.)

Otherwise, if the work is a derivative of the Library, you may distribute the object code for the work under the terms of Section 6. Any executables containing that work also fall under Section 6, whether or not they are linked directly with the Library itself.

6. As an exception to the Sections above, you may also compile or link a "work that uses the Library" with the Library to produce a work containing portions of the Library, and distribute that work under terms of your choice, provided that the terms permit modification of the work for the customer's own use and reverse engineering for debugging such modifications.

You must give prominent notice with each copy of the work that the Library is used in it and that the Library and its use are covered by this License. You must supply a copy of this License. If the work during execution displays copyright notices, you must include the copyright notice for the Library among them, as well as a reference directing the user to the copy of this License. Also, you must do one of these things:

- a) Accompany the work with the complete corresponding machine-readable source code for the Library including whatever changes were used in the work (which must be distributed under Sections 1 and 2 above); and, if the work is an executable linked with the Library, with the complete machine-readable "work that uses the Library", as object code and/or source code, so that the user can modify the Library and then relink to produce a modified executable containing the modified Library. (It is understood that the user who changes the contents of definitions files in the Library will not necessarily be able to recompile the application to use the modified definitions.)
- b) Accompany the work with a written offer, valid for at least three years, to give the same user the materials specified in Subsection 6a, above, for a charge no more than the cost of performing this distribution.
- c) If distribution of the work is made by offering access to copy from a designated place, offer equivalent access to copy the above specified materials from the same place.
- d) Verify that the user has already received a copy of these

materials or that you have already sent this user a copy.

For an executable, the required form of the "work that uses the Library" must include any data and utility programs needed for reproducing the executable from it. However, as a special exception, the source code distributed need not include anything that is normally distributed (in either source or binary form) with the major components (compiler, kernel, and so on) of the operating system on which the executable runs, unless that component itself accompanies the executable.

It may happen that this requirement contradicts the license restrictions of other proprietary libraries that do not normally accompany the operating system. Such a contradiction means you cannot use both them and the Library together in an executable that you distribute.

7. You may place library facilities that are a work based on the Library side-by-side in a single library together with other library facilities not covered by this License, and distribute such a combined library, provided that the separate distribution of the work based on the Library and of the other library facilities is otherwise permitted, and provided that you do these two things:

- a) Accompany the combined library with a copy of the same work based on the Library, uncombined with any other library facilities. This must be distributed under the terms of the Sections above.
- b) Give prominent notice with the combined library of the fact that part of it is a work based on the Library, and explaining where to find the accompanying uncombined form of the same work.

8. You may not copy, modify, sublicense, link with, or distribute the Library except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, link with, or distribute the Library is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

9. You are not required to accept this License, since you have not signed it. However, nothing else grants you permission to modify or distribute the Library or its derivative works. These actions are prohibited by law if you do not accept this License. Therefore, by modifying or distributing the Library (or any work based on the Library), you indicate your acceptance of this License to do so, and all its terms and conditions for copying, distributing or modifying the Library or works based on it.

10. Each time you redistribute the Library (or any work based on the Library), the recipient automatically receives a license from the

original licensor to copy, distribute, link with or modify the Library subject to these terms and conditions. You may not impose any further restrictions on the recipients' exercise of the rights granted herein. You are not responsible for enforcing compliance by third parties to this License.

11. If, as a consequence of a court judgment or allegation of patent infringement or for any other reason (not limited to patent issues), conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot distribute so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not distribute the Library at all. For example, if a patent license would not permit royalty-free redistribution of the Library by all those who receive copies directly or indirectly through you, then the only way you could satisfy both it and this License would be to refrain entirely from distribution of the Library.

If any portion of this section is held invalid or unenforceable under any particular circumstance, the balance of the section is intended to apply, and the section as a whole is intended to apply in other circumstances.

It is not the purpose of this section to induce you to infringe any patents or other property right claims or to contest validity of any such claims; this section has the sole purpose of protecting the integrity of the free software distribution system which is implemented by public license practices. Many people have made generous contributions to the wide range of software distributed through that system in reliance on consistent application of that system; it is up to the author/donor to decide if he or she is willing to distribute software through any other system and a licensee cannot impose that choice.

This section is intended to make thoroughly clear what is believed to be a consequence of the rest of this License.

12. If the distribution and/or use of the Library is restricted in certain countries either by patents or by copyrighted interfaces, the original copyright holder who places the Library under this License may add an explicit geographical distribution limitation excluding those countries, so that distribution is permitted only in or among countries not thus excluded. In such case, this License incorporates the limitation as if written in the body of this License.

13. The Free Software Foundation may publish revised and/or new versions of the Library General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Library specifies a version number of this License which applies to it and

"any later version", you have the option of following the terms and conditions either of that version or of any later version published by the Free Software Foundation. If the Library does not specify a license version number, you may choose any version ever published by the Free Software Foundation.

14. If you wish to incorporate parts of the Library into other free programs whose distribution conditions are incompatible with these, write to the author to ask for permission. For software which is copyrighted by the Free Software Foundation, write to the Free Software Foundation; we sometimes make exceptions for this. Our decision will be guided by the two goals of preserving the free status of all derivatives of our free software and of promoting the sharing and reuse of software generally.

NO WARRANTY

15. BECAUSE THE LIBRARY IS LICENSED FREE OF CHARGE, THERE IS NO WARRANTY FOR THE LIBRARY, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE LIBRARY "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE LIBRARY IS WITH YOU. SHOULD THE LIBRARY PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MAY MODIFY AND/OR REDISTRIBUTE THE LIBRARY AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE LIBRARY (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE LIBRARY TO OPERATE WITH ANY OTHER SOFTWARE), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

END OF TERMS AND CONDITIONS

Appendix: How to Apply These Terms to Your New Libraries

If you develop a new library, and you want it to be of the greatest possible use to the public, we recommend making it free software that everyone can redistribute and change. You can do so by permitting redistribution under these terms (or, alternatively, under the terms of the ordinary General Public License).

To apply these terms, attach the following notices to the library. It is safest to attach them to the start of each source file to most effectively convey the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full

notice is found.

```
<one line to give the library's name and a brief idea of what it does.>
Copyright (C) <year> <name of author>
```

This library is free software; you can redistribute it and/or modify it under the terms of the GNU Library General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This library is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Library General Public License for more details.

You should have received a copy of the GNU Library General Public License along with this library; if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA

Also add information on how to contact you by electronic and paper mail.

You should also get your employer (if you work as a programmer) or your school, if any, to sign a "copyright disclaimer" for the library, if necessary. Here is a sample; alter the names:

Yoyodyne, Inc., hereby disclaims all copyright interest in the library 'Frob' (a library for tweaking knobs) written by James Random Hacker.

```
<signature of Ty Coon>, 1 April 1990
Ty Coon, President of Vice
```

That's all there is to it!

5.6.23 Netscape Public License

Some third-party software packages distributed with the Search Appliance are governed by the Netscape Public License, reproduced below. See the Third-Party Software section, p. 121, for a list of applicable packages.

Netscape Public License version 1.1

AMENDMENTS The Netscape Public License Version 1.1 ("NPL") consists of the Mozilla Public License Version 1.1 with the following Amendments, including Exhibit A-Netscape Public License. Files identified with "Exhibit A-Netscape Public License" are governed by the Netscape Public License Version 1.1.

Additional Terms applicable to the Netscape Public License.

I. Effect.

These additional terms described in this Netscape Public License – Amendments shall apply to the Mozilla

Communicator client code and to all Covered Code under this License.

II. "Netscape's Branded Code" means Covered Code that Netscape distributes and/or permits others to distribute under one or more trademark(s) which are controlled by Netscape but which are not licensed for use under this License.

III. Netscape and logo. This License does not grant any rights to use the trademarks "Netscape", the "Netscape N and horizon" logo or the "Netscape lighthouse" logo, "Netcenter", "Gecko", "Java" or "JavaScript", "Smart Browsing" even if such marks are included in the Original Code or Modifications.

IV. Inability to Comply Due to Contractual Obligation. Prior to licensing the Original Code under this License, Netscape has licensed third party code for use in Netscape's Branded Code. To the extent that Netscape is limited contractually from making such third party code available under this License, Netscape may choose to reintegrate such code into Covered Code without being required to distribute such code in Source Code form, even if such code would otherwise be considered "Modifications" under this License.

V. Use of Modifications and Covered Code by Initial Developer.

V.1. In General. The obligations of Section 3 apply to Netscape, except to the extent specified in this Amendment, Section V.2 and V.3.

V.2. Other Products. Netscape may include Covered Code in products other than the Netscape's Branded Code which are released by Netscape during the two (2) years following the release date of the Original Code, without such additional products becoming subject to the terms of this License, and may license such additional products on different terms from those contained in this License.

V.3. Alternative Licensing. Netscape may license the Source Code of Netscape's Branded Code, including Modifications incorporated therein, without such Netscape Branded Code becoming subject to the terms of this License, and may license such Netscape Branded Code on different terms from those contained in this License.

VI. Litigation. Notwithstanding the limitations of Section 11 above, the provisions regarding litigation in Section 11(a), (b) and (c) of the License shall apply to all disputes relating to this License.

EXHIBIT A-Netscape Public License.

"The contents of this file are subject to the Netscape Public License Version 1.1 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.mozilla.org/NPL/> Software distributed under the License is distributed on an "AS IS" basis, WITHOUT WARRANTY OF ANY KIND, either express or implied. See the License for the specific language governing rights and limitations under the License.

The Original Code is Mozilla Communicator client code, released March 31, 1998.

The Initial Developer of the Original Code is Netscape Communications Corporation. Portions created by Netscape are Copyright (C) 1998-1999 Netscape Communications Corporation. All Rights Reserved.
Contributor(s): _____.

Alternatively, the contents of this file may be used under the terms of the — license (the "[—] License"), in which case the provisions of [—] License are applicable instead of those above. If you wish to allow use of your version of this file only under the terms of the [—] License and not to allow others to use your version of this file under the NPL, indicate your decision by deleting the provisions above and replace them

with the notice and other provisions required by the [—] License. If you do not delete the provisions above, a recipient may use your version of this file under either the NPL or the [—] License.”

MOZILLA PUBLIC LICENSE

Version 1.1

1. Definitions.

1.0.1. "Commercial Use" means distribution or otherwise making the Covered Code available to a third party.

1.1. "Contributor" means each entity that creates or contributes to the creation of Modifications.

1.2. "Contributor Version" means the combination of the Original Code, prior Modifications used by a Contributor, and the Modifications

made by that particular Contributor.

1.3. "Covered Code" means the Original Code or Modifications or the combination of the Original Code and Modifications, in each case including portions thereof.

1.4. "Electronic Distribution Mechanism" means a mechanism generally accepted in the software development community for the electronic transfer of data.

1.5. "Executable" means Covered Code in any form other than Source Code.

1.6. "Initial Developer" means the individual or entity identified as the Initial Developer in the Source Code notice required by **Exhibit**

A.

1.7. "Larger Work" means a work which combines Covered Code or portions thereof with code not governed by the terms of this License.

1.8. "License" means this document.

1.8.1. "Licensable" means having the right to grant, to the maximum extent possible, whether at the time of the initial grant or subsequently acquired, any and all of the rights conveyed herein.

1.9. "Modifications" means any addition to or deletion from the substance or structure of either the Original Code or any previous Modifications. When Covered Code is released as a series of files, a Modification is:

A. Any addition to or deletion from the contents of a file containing Original Code or previous Modifications.

B. Any new file that contains any part of the Original Code or previous Modifications.

1.10. "Original Code" means Source Code of computer software code which is described in the Source Code notice required by **Exhibit A** as Original Code, and which, at the time of its release under this License is not already Covered Code governed by this License.

1.10.1. "Patent Claims" means any patent claim(s), now owned or hereafter acquired, including without limitation, method, process, and apparatus claims, in any patent Licensable by grantor.

1.11. "Source Code" means the preferred form of the Covered Code for making modifications to it,

including all modules it contains, plus any associated interface definition files, scripts used to control compilation and installation of an Executable, or source code differential comparisons against either the Original Code or another well known, available Covered Code of the Contributor's choice. The Source Code can be in a compressed or archival form, provided the appropriate decompression or de-archiving software is widely available for no charge.

1.12. "You" (or "Your") means an individual or a legal entity exercising rights under, and complying with all of the terms of, this License or a future version of this License issued under Section 6.1. For legal entities, "You" includes any entity which controls, is controlled by, or is under common control with You. For purposes of this definition, "control" means (a) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (b) ownership of more than fifty percent (50) beneficial ownership of such entity.

2. Source Code License.

2.1. The Initial Developer Grant.

The Initial Developer hereby grants You a world-wide, royalty-free, non-exclusive license, subject to third party intellectual property claims:

- (a) under intellectual property rights (other than patent or trademark) Licensable by Initial Developer to use, reproduce, modify, display, perform, sublicense and distribute the Original Code (or portions thereof) with or without Modifications, and/or as part of a Larger Work; and
- (b) under Patents Claims infringed by the making, using or selling of Original Code, to make, have made, use, practice, sell, and offer for sale, and/or otherwise dispose of the Original Code (or portions thereof).
- (c) the licenses granted in this Section 2.1(a) and (b) are effective on the date Initial Developer first distributes Original Code under the terms of this License.
- (d) Notwithstanding Section 2.1(b) above, no patent license is granted: 1) for code that You delete from the Original Code; 2) separate from the Original Code; or 3) for infringements caused by: i) the modification of the Original Code or ii) the combination of the Original Code with other software or devices.

2.2. Contributor Grant.

Subject to third party intellectual property claims, each Contributor hereby grants You a world-wide, royalty-free, non-exclusive license

- (a) under intellectual property rights (other than patent or trademark) Licensable by Contributor, to use, reproduce, modify, display, perform, sublicense and distribute the Modifications created by such Contributor (or portions thereof) either on an unmodified basis, with other Modifications, as Covered Code and/or as part of a Larger Work; and
- (b) under Patent Claims infringed by the making, using, or selling of Modifications made by that Contributor either alone and/or in combination with its Contributor Version (or portions of such combination), to make, use, sell, offer for sale, have made, and/or otherwise dispose of: 1) Modifications made by that Contributor (or portions thereof); and 2) the combination of Modifications made by that Contributor with its Contributor Version (or portions of such combination).
- (c) the licenses granted in Sections 2.2(a) and 2.2(b) are effective on the date Contributor first makes Commercial Use of the Covered Code.

(d) Notwithstanding Section 2.2(b) above, no patent license is granted: 1) for any code that Contributor has deleted from the Contributor Version; 2) separate from the Contributor Version; 3) for infringements caused by: i) third party modifications of Contributor Version or ii) the combination of Modifications made by that Contributor with other software (except as part of the Contributor Version) or other devices; or 4) under Patent Claims infringed by Covered Code in the absence of Modifications made by that Contributor.

3. Distribution Obligations.

3.1. Application of License.

The Modifications which You create or to which You contribute are governed by the terms of this License, including without limitation Section 2.2. The Source Code version of Covered Code may be distributed only under the terms of this License or a future version of this License released under Section 6.1, and You must include a copy of this License with every copy of the Source Code You distribute. You may not offer or impose any terms on any Source Code version that alters or restricts the applicable version of this License or the recipients' rights hereunder. However, You may include an additional document offering the additional rights described in Section 3.5.

3.2. Availability of Source Code.

Any Modification which You create or to which You contribute must be made available in Source Code form under the terms of this License either on the same media as an Executable version or via an accepted Electronic Distribution Mechanism to anyone to whom you made an Executable version available; and if made available via Electronic Distribution Mechanism, must remain available for at least twelve (12) months after the date it initially became available, or at least six (6) months after a subsequent version of that particular Modification has been made available to such recipients. You are responsible for ensuring that the Source Code version remains available even if the Electronic Distribution Mechanism is maintained by a third party.

3.3. Description of Modifications.

You must cause all Covered Code to which You contribute to contain a file documenting the changes You made to create that Covered Code and the date of any change. You must include a prominent statement that the Modification is derived, directly or indirectly, from Original Code provided by the Initial Developer and including the name of the Initial Developer in (a) the Source Code, and (b) in any notice in an Executable version or related documentation in which You describe the origin or ownership of the Covered Code.

3.4. Intellectual Property Matters

(a) Third Party Claims.

If Contributor has knowledge that a license under a third party's intellectual property rights is required to exercise the rights granted by such Contributor under Sections 2.1 or 2.2, Contributor must include a text file with the Source Code distribution titled "LEGAL" which describes the claim and the party making the claim in sufficient detail that a recipient will know whom to contact. If Contributor obtains such knowledge after the Modification is made available as described in Section 3.2, Contributor shall promptly modify the LEGAL file in all copies Contributor makes available thereafter and shall take other steps (such as notifying appropriate mailing lists or newsgroups) reasonably calculated to inform those who received the Covered Code that new knowledge has been obtained.

(b) Contributor APIs.

If Contributor's Modifications include an application programming interface and Contributor has knowledge of patent licenses which are reasonably necessary to implement that API, Contributor must also include this information in the LEGAL file.

(c) Representations.

Contributor represents that, except as disclosed pursuant to Section 3.4(a) above, Contributor believes that Contributor's Modifications are Contributor's original creation(s) and/or Contributor has sufficient rights to grant the rights conveyed by this License.

3.5. Required Notices.

You must duplicate the notice in **Exhibit A** in each file of the Source Code. If it is not possible to put such notice in a particular Source Code file due to its structure, then You must include such notice in a location (such as a relevant directory) where a user would be likely to look for such a notice. If You created one or more Modification(s) You may add your name as a Contributor to the notice described in **Exhibit A**. You must also duplicate this License in any documentation for the Source Code where You describe recipients' rights or ownership rights relating to Covered Code. You may choose to offer, and to charge a fee for, warranty, support, indemnity or liability obligations to one or more recipients of Covered Code. However, You may do so only on Your own behalf, and not on behalf of the Initial Developer or any Contributor. You must make it absolutely clear than any such warranty, support, indemnity or liability obligation is offered by You alone, and You hereby agree to indemnify the Initial Developer and every Contributor for any liability incurred by the Initial Developer or such Contributor as a result of warranty, support, indemnity or liability terms You offer.

3.6. Distribution of Executable Versions.

You may distribute Covered Code in Executable form only if the requirements of Section **3.1-3.5** have been met for that Covered Code, and if You include a notice stating that the Source Code version of the Covered Code is available under the terms of this License, including a description of how and where You have fulfilled the obligations of Section **3.2**. The notice must be conspicuously included in any notice in an Executable version, related documentation or collateral in which You describe recipients' rights relating to the Covered Code. You may distribute the Executable version of Covered Code or ownership rights under a license of Your choice, which may contain terms different from this License, provided that You are in compliance with the terms of this License and that the license for the Executable version does not attempt to limit or alter the recipient's rights in the Source Code version from the rights set forth in this License. If You distribute the Executable version under a different license You must make it absolutely clear that any terms which differ from this License are offered by You alone, not by the Initial Developer or any Contributor. You hereby agree to indemnify the Initial Developer and every Contributor for any liability incurred by the Initial Developer or such Contributor as a result of any such terms You offer.

3.7. Larger Works.

You may create a Larger Work by combining Covered Code with other code not governed by the terms of this License and distribute the Larger Work as a single product. In such a case, You must make sure the requirements of this License are fulfilled for the Covered Code.

4. Inability to Comply Due to Statute or Regulation.

If it is impossible for You to comply with any of the terms of this License with respect to some or all of the Covered Code due to statute, judicial order, or regulation then You must: (a) comply with the terms of this

License to the maximum extent possible; and (b) describe the limitations and the code they affect. Such description must be included in the LEGAL file described in Section 3.4 and must be included with all distributions of the Source Code. Except to the extent prohibited by statute or regulation, such description must be sufficiently detailed for a recipient of ordinary skill to be able to understand it.

5. Application of this License.

This License applies to code to which the Initial Developer has attached the notice in **Exhibit A** and to related Covered Code.

6. Versions of the License.

6.1. New Versions.

Netscape Communications Corporation ("Netscape") may publish revised and/or new versions of the License from time to time. Each version will be given a distinguishing version number.

6.2. Effect of New Versions.

Once Covered Code has been published under a particular version of the License, You may always continue to use it under the terms of that version. You may also choose to use such Covered Code under the terms of any subsequent version of the License published by Netscape. No one other than Netscape has the right to modify the terms applicable to Covered Code created under this License.

6.3. Derivative Works.

If You create or use a modified version of this License (which you may only do in order to apply it to code which is not already Covered Code governed by this License), You must (a) rename Your license so that the phrases "Mozilla", "MOZILLAPL", "MOZPL", "Netscape", "MPL", "NPL" or any confusingly similar phrase do not appear in your license (except to note that your license differs from this License) and (b) otherwise make it clear that Your version of the license contains terms which differ from the Mozilla Public License and Netscape Public License. (Filling in the name of the Initial Developer, Original Code or Contributor in the notice described in **Exhibit A** shall not of themselves be deemed to be modifications of this License.)

7. DISCLAIMER OF WARRANTY.

COVERED CODE IS PROVIDED UNDER THIS LICENSE ON AN "AS IS" BASIS, WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, WITHOUT LIMITATION, WARRANTIES THAT THE COVERED CODE IS FREE OF DEFECTS, MERCHANTABLE, FIT FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE COVERED CODE IS WITH YOU. SHOULD ANY COVERED CODE PROVE DEFECTIVE IN ANY RESPECT, YOU (NOT THE INITIAL DEVELOPER OR ANY OTHER CONTRIBUTOR) ASSUME THE COST OF ANY NECESSARY SERVICING, REPAIR OR CORRECTION. THIS DISCLAIMER OF WARRANTY CONSTITUTES AN ESSENTIAL PART OF THIS LICENSE. NO USE OF ANY COVERED CODE IS AUTHORIZED HEREUNDER EXCEPT UNDER THIS DISCLAIMER.

8. TERMINATION.

8.1. This License and the rights granted hereunder will terminate automatically if You fail to comply with terms herein and fail to cure such breach within 30 days of becoming aware of the breach. All sublicenses to

the Covered Code which are properly granted shall survive any termination of this License. Provisions which, by their nature, must remain in effect beyond the termination of this License shall survive.

8.2. If You initiate litigation by asserting a patent infringement claim (excluding declaratory judgment actions) against Initial Developer or a Contributor (the Initial Developer or Contributor against whom You file such action is referred to as "Participant") alleging that:

(a) such Participant's Contributor Version directly or indirectly infringes any patent, then any and all rights granted by such Participant to You under Sections 2.1 and/or 2.2 of this License shall, upon 60 days notice from Participant terminate prospectively, unless if within 60 days after receipt of notice You either: (i) agree in writing to pay Participant a mutually agreeable reasonable royalty for Your past and future use of Modifications made by such Participant, or (ii) withdraw Your litigation claim with respect to the Contributor Version against such Participant. If within 60 days of notice, a reasonable royalty and payment arrangement are not mutually agreed upon in writing by the parties or the litigation claim is not withdrawn, the rights granted by Participant to You under Sections 2.1 and/or 2.2 automatically terminate at the expiration of the 60 day notice period specified above.

(b) any software, hardware, or device, other than such Participant's Contributor Version, directly or indirectly infringes any patent, then any rights granted to You by such Participant under Sections 2.1(b) and 2.2(b) are revoked effective as of the date You first made, used, sold, distributed, or had made, Modifications made by that Participant.

8.3. If You assert a patent infringement claim against Participant alleging that such Participant's Contributor Version directly or indirectly infringes any patent where such claim is resolved (such as by license or settlement) prior to the initiation of patent infringement litigation, then the reasonable value of the licenses granted by such Participant under Sections 2.1 or 2.2 shall be taken into account in determining the amount or value of any payment or license.

8.4. In the event of termination under Sections 8.1 or 8.2 above, all end user license agreements (excluding distributors and resellers) which have been validly granted by You or any distributor hereunder prior to termination shall survive termination.

9. LIMITATION OF LIABILITY.

UNDER NO CIRCUMSTANCES AND UNDER NO LEGAL THEORY, WHETHER TORT (INCLUDING NEGLIGENCE), CONTRACT, OR OTHERWISE, SHALL YOU, THE INITIAL DEVELOPER, ANY OTHER CONTRIBUTOR, OR ANY DISTRIBUTOR OF COVERED CODE, OR ANY SUPPLIER OF ANY OF SUCH PARTIES, BE LIABLE TO ANY PERSON FOR ANY INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF GOODWILL, WORK STOPPAGE, COMPUTER FAILURE OR MALFUNCTION, OR ANY AND ALL OTHER COMMERCIAL DAMAGES OR LOSSES, EVEN IF SUCH PARTY SHALL HAVE BEEN INFORMED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION OF LIABILITY SHALL NOT APPLY TO LIABILITY FOR DEATH OR PERSONAL INJURY RESULTING FROM SUCH PARTY'S NEGLIGENCE TO THE EXTENT APPLICABLE LAW PROHIBITS SUCH LIMITATION. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OR LIMITATION OF INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THIS EXCLUSION AND LIMITATION MAY NOT APPLY TO YOU.

10. U.S. GOVERNMENT END USERS.

The Covered Code is a "commercial item," as that term is defined in 48 C.F.R. 2.101 (Oct. 1995), consisting of "commercial computer software" and "commercial computer software documentation," as such terms are used in 48 C.F.R. 12.212 (Sept. 1995). Consistent with 48 C.F.R. 12.212 and 48 C.F.R. 227.7202-1 through 227.7202-4 (June 1995), all U.S. Government End Users acquire Covered Code with only those rights set forth herein.

11. MISCELLANEOUS.

This License represents the complete agreement concerning subject matter hereof. If any provision of this License is held to be unenforceable, such provision shall be reformed only to the extent necessary to make it enforceable. This License shall be governed by California law provisions (except to the extent applicable law, if any, provides otherwise), excluding its conflict-of-law provisions. With respect to disputes in which at least one party is a citizen of, or an entity chartered or registered to do business in the United States of America, any litigation relating to this License shall be subject to the jurisdiction of the Federal Courts of the Northern District of California, with venue lying in Santa Clara County, California, with the losing party responsible for costs, including without limitation, court costs and reasonable attorneys' fees and expenses. The application of the United Nations Convention on Contracts for the International Sale of Goods is expressly excluded. Any law or regulation which provides that the language of a contract shall be construed against the drafter shall not apply to this License.

12. RESPONSIBILITY FOR CLAIMS.

As between Initial Developer and the Contributors, each party is responsible for claims and damages arising, directly or indirectly, out of its utilization of rights under this License and You agree to work with Initial Developer and Contributors to distribute such responsibility on an equitable basis. Nothing herein is intended or shall be deemed to constitute any admission of liability.

13. MULTIPLE-LICENSED CODE.

Initial Developer may designate portions of the Covered Code as "Multiple-Licensed". "Multiple-Licensed" means that the Initial Developer permits you to utilize portions of the Covered Code under Your choice of the NPL or the alternative licenses, if any, specified by the Initial Developer in the file described in Exhibit A.

EXHIBIT A -Mozilla Public License.

"The contents of this file are subject to the Mozilla Public License Version 1.1 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.mozilla.org/MPL/> Software distributed under the License is distributed on an "AS IS" basis, WITHOUT WARRANTY OF ANY KIND, either express or implied. See the License for the specific language governing rights and limitations under the License. The Original Code is _____ . The Initial Developer of the Original Code is _____. Portions created by _____ are Copyright (C) _____. All Rights Reserved. Contributor(s): _____. Alternatively, the contents of this file may be used under the terms of the _____ license (the "[] License"), in which case the provisions of [] License are applicable instead of those above. If you wish to allow use of your version of this file only under the terms of the [] License and not to allow others to use your version of this file under the MPL, indicate your decision by deleting the provisions above and replace them with the notice and other provisions required by the [] License. If you do not delete the provisions above, a recipient may use your version of this file under either the MPL or the [] License."

[NOTE: The text of this Exhibit A may differ slightly from the text of the notices in the Source Code files of the Original Code. You should use the text of this Exhibit A rather than the text found in the Original Code Source Code for Your Modifications.]

5.7 XML Elements in Search Results

Search results can be sent as XML from the Search Appliance to the host server. This section describes the XML elements. The elements in the following table are listed in the approximate order that they are sent.

Table 5.7: Search Results XML Elements

Element	Description
<code>¿?xml version="1.0"?¿</code>	The version of this xml
<code>¿ThunderstoneResults¿</code>	Tag that encloses all of results from appliance
<code>¿Query¿</code>	Search string that was submitted to appliance
<code>¿ResultsPerSiteQuery¿</code>	User's max-results-per-site number
<code>¿TextQuery¿</code>	The text part of the query (sans <code>site:host</code>)
<code>¿SiteQuery¿</code>	The site query (from <code>site:host</code> or <code>sq var</code>)
<code>¿UrlRoot¿</code>	The URL root of the search script
<code>¿Profile¿</code>	The appliance profile, which is a set of walk and search settings
<code>¿dropXSL¿</code>	If yes removes XSL from results
<code>¿authUser¿</code>	The user that was authenticated via the Proxy Module.
<code>¿Summary¿</code>	Encloses search results summary. The <code>¿result¿</code> section follows.
<code>¿Start¿</code>	First result item to list
<code>¿End¿</code>	Last result item to list
<code>¿TotalNum¿</code>	Total number of result items
<code>¿Total¿</code>	Text string that represents the total number of result items
<code>¿CurOrder¿</code>	Text that describes the order by which results are listed
<code>¿OrderLink¿</code>	Link that provides an alternative sorting order results list
<code>¿OrderType¿</code>	Text that describes OrderLink
<code>¿FirstPage¿</code>	Page number of page to show first
<code>¿Credit¿</code>	Text to introduce credit image
<code>¿CreditImage¿</code>	The URL of credit image
<code>¿Pages¿</code>	Tag that groups tags for a specific page of results
<code>¿PageLink¿</code>	Link to a certain page of the results
<code>¿PageNumber¿</code>	Page number a page of results
<code>¿NextLink¿</code>	Link to the next page of results (current page plus 1)
<code>¿LastPage¿</code>	Link to the previous page of results (current page minus 1)
<code>¿Result¿</code>	Tag that contains all results elements
<code>¿Profile¿</code>	Name of the profile, which holds settings related to this search
<code>¿Num¿</code>	Number of this result item
<code>¿Id¿</code>	Identifier for this result item
<code>¿ResultTitle¿</code>	Title of the page of this result item
<code>¿Url¿</code>	The URL of the page for this result item
<code>¿UrlPDFHi¿</code>	The URL to highlight this PDF item in user's Acrobat Viewer
<code>¿UrlDisplay¿</code>	Displayed URL for this result item
<code>¿RawRank¿</code>	The raw relevance rank value for this result item (0-1000)
<code>¿ScaledRank¿</code>	Raw rank scaled up for a more-like-this search (0-1000)
<code>¿PercentRank¿</code>	ScaledRank as a percentage (0-100)
<code>¿DocSize¿</code>	Size (bytes) of the page of this result item
<code>¿Depth¿</code>	Number of links walked from Base URL to this URL
<code>¿UrlSimilar¿</code>	The URL to search for pages similar to this result item
<code>¿UrlInfo¿</code>	The URL for context of answers within a matching document
<code>¿UrlParents¿</code>	The URL of pages that link to the page of this search result item
<code>¿Modified¿</code>	Date and time that the page of this result item was last modified
<code>¿Abstract¿</code>	Brief text surrounding the matched word or phrase
<code>¿Charset¿</code>	Character set of the page
<code>¿SiteName¿</code>	The name of the site for this result item
<code>¿UrlMoreResultsFromSite¿</code>	The URL for more results from this site

Chapter 6

Search Interface Help

6.1 Forming a Query

The Search Appliance's search can be as simple or as complex as you need it to be. Usually you will just need to enter a few words that best describe that which you are trying to locate. To perform more complicated searches you might use any combination of logic operators, special pattern matchers, concept expansion, or proximity operations.

Example: nature conservation organization

6.1.1 Query Rules of Thumb

- If you get too many junk or nonsense answers, try:
 - Add some more words to your query.
 - Decrease the range of the Proximity control.
 - Change the Word Forms control to Exact.
 - Look at the Match Info and see why they are showing up.
 - Use the Exclusion Operator (-) to remove unwanted terms.
 - If you are searching for a phrase, hyphenate the words together.
- If you don't get any answers, or just too few:
 - Remove some more words to your query.
 - Examine your spelling.
 - Increase the scope of the Proximity control.
 - It just might not be there?

6.1.2 Overview of Query Abilities

The Search Appliance is based on Taxis and as such it shares its text query abilities with all of Thunderstone's products. Throughout our documentation you will see references to Metamorph or Taxis. This is because all of our products share a common text query language. This document provides only a brief overview of this language.

If you'd like to know more see the online manual at
http://www.thunderstone.com/site/texisman/link_mmq.html.

6.1.3 Controlling Proximity

Mastering the usage of proximity gives the ability to locate answers with greater precision. The Search Appliance input form gives you several options to control the search proximity:

line All query terms must occur on the same line

sentence Query items should all reside within the same sentence

paragraph Within the same paragraph or text block

page All items must occur within same HTML document (the default)

A bar-graph display will be shown any time a ranking search was performed (eg. all searches except Show Parents).

6.1.4 Ranking Factors

The ranking algorithm takes into consideration relative word ordering, word proximity, database frequency, document frequency, and position in text. The relative importance of these factors in computing the quality of a hit can be altered under RANKING FACTORS on the Options page.

6.1.5 Keywords Phrases and Wild-cards

To locate words, just type them in as you would in a word processor. Letter cases will be ignored.

The wild-card character * (asterisk) may be used to match just the prefix of a word or to ignore the middle of something.

If the item you wish to locate is more complicated than the simple * wild-card can accomplish, try using the regular expression matcher (<http://www.thunderstone.com/taxis/site/pages/regexp.html>).

To locate a number of adjacent words in a specific order, surround them with " (double quotation) characters. Putting a - (hyphen) between words will also force order and one word proximity.

* see Word Forms (6.2, p. 181)

Table 6.1: Query examples

Query	Locates
john	john, John
"john public"	John Public
web-browser	Web browser, web-browser
John*Public	John Q. Public, John Public
456*a*def	1-456-789-ABCDEF
activate	activate, activation, activated, ... *

6.1.6 Applying Search Logic

Taxis and Metamorph use set logic for text queries. Set logic is easier to use and provides more abilities than boolean. The examples below make reference to single keywords, but keep in mind that each keyword can represent an entire list of things or any of the special pattern matchers.

Sets (or lists) of things are specified by placing the elements within parenthesis, separated by commas.

Example: *(bob,joe,sam,sue)* . In the examples below, you could replace any of the keywords with a list like this.

The default behavior of the search is to locate an intersection (or 'AND') of every element within a query. This means that the query: *"microsoft bob interface"* is the equivalent to the boolean query: *"microsoft AND bob AND interface"* .

- (without) The - (minus) is the most commonly used logic symbol. It means the answer should EXCLUDE references to that item.

+ (mandatory) The + (plus) symbol in front of a search item means that the answer MUST INCLUDE that item. This is generally used in conjunction with the permutation operation.

@N (permute) The @ followed by a number indicates how many intersections to locate of the terms in your query. This may be confusing at first, but it is very powerful.

Table 6.2: Search Logic Examples

Query	Finds
bob sam joe	Bob with Sam and Joe
bob sam -joe	Bob with Sam without Joe
bob sam joe @1	Bob with Sam, or Bob with Joe, or Joe with Sam
A B C D @1	AB or AC or AD or BC or BD or CD
+A B C D @1	ABC or ABD or ACD
A B C -D @1	(AB or AC or BC) without D

The plus(+) and minus(-) operators must be attached to the term to which they apply. There must be a space between the operator and any preceding term.

Correct	Incorrect
bob +sam -joe	bob + sam - joe
	bob+sam-joe

6.1.7 Natural Language Query

You may enter a query in the form of a sentence or question. The software will automatically identify the important words and phrases within your query and remove the “noise words”.

Example: What is the state of the art in text retrieval?

The software will search for: state of the art AND text AND retrieval

6.1.8 Using the Special Pattern Matchers

These pattern matchers are used to locate hard-to-find items within text:

- Regular expression matching for complex patterns
<http://www.thunderstone.com/taxis/site/pages/regexp.html>
- Approximate pattern matching for fuzzy searches
<http://www.thunderstone.com/taxis/site/pages/xpm.html>
- Numeric pattern matching for finding quantities
<http://www.thunderstone.com/taxis/site/pages/npm.html>

If improperly used these pattern matchers can slow queries. Therefore they require other keyword(s) in the query and are disabled entirely under Page proximity. For more details see the Vortex manual on Query Protection (http://www.thunderstone.com/site/vortexman/link_qprot.html).

Table 6.3: Pattern Matcher Examples

Query	Matcher	Finds
ronald %regan	Approx	Ronald Raygun, Ronald Re-an, Ronald 8eagan
%75MYPARTNO9045d/6a	Approx	Anything within 75% of looking like MYPARTNO9045d/6a
/19[789][0-9]	RegEXpr	1970-1999
/[1-9]{3}\-[0-9]{4}	RegEXpr	Phone numbers: 555-1212, 820-2200
#87	Numeric	four score and seven, 87
#>0<1	Numeric	Fractions like 9/16, 55%, 0.123, 15 nanoseconds

Table 6.4: Word Form Examples

Word	president
EXACT	president
PLURAL	(above) + presidents president's
ANY	(above) + presidential presidency preside presides presiding presided
Word	tight
EXACT	tight
PLURAL	(above) + tights
ANY	(above) + tightly tightening tightened tighter tightest
Word	program
EXACT	program
PLURAL	(above) + programs program's
ANY	(above) + programming programmatic programmed programmer programmable

6.1.9 Invoking Thesaurus Expansion

The Search Appliance has a vocabulary of over 250,000 word and phrase associations. Each entry is generally classifiable by either its meaning or part of speech.

Depending on the administrator's Synonyms setting for this profile, synonyms may already be included for each term in your query. If not, synonyms may be included for individual terms within your query by preceding them with a ~ (tilde) character.

6.2 Using Word Forms

The `Word forms` options give you control over how many variations of your query terms will be sought in your search.

Exact: Only exact matches will be allowed. (the default)

Plural & possessives: Plural and possessive forms will be found. (s, es, 's)

Any word forms: As many word forms as can be derived will be located.

We call this morpheme processing, and it is generally smarter than a traditional “stemming” algorithm. It doesn't just rip the end off a word, it actually checks to see if it could be a valid form of the search term.

More information is available at

http://www.thunderstone.com/site/texisman/link_ling.html.

Notes: Thesaurus terms are also treated in the same manner. Words smaller than 4-5 characters will not be morpheme processed.

6.3 Controlling Proximity

These options give you control over the region in which a match must be found.

line: match terms must be located within the same line.

sentence: all terms within the same sentence.

paragraph: match terms must be located within the same paragraph.

page: (default) all terms within the same document.

In all cases the best possible matches for your query are located and ordered by decreasing quality. A bar graph is produced to indicate the quality of each answer.

6.4 Interpreting Search Results

Note: *The look and feel described here is for the standard search interface. The interface may have been customized by the web site administrator.*

When a query is submitted it will come back with another query form and up to 10 matching documents. If there are more than 10 answers, a link at the top and bottom of the list will allow you to view the next 10 in sequence.

The input form at the top allows you to further tailor your query to home-in on the desired answers, or to submit a completely new query without having to navigate back to the original input form.

Each answer in the result set will have a format similar to the following:

1: THE DOCUMENT TITLE (hyperlink to original)	84%***** ____
This is the document abstract. It consists	Size: 11K
of the text around the first hit within the	Depth: 3
matching document...	Find Similar
http://www.thesite.com/thepage.html	Match Info
	Show Parents

The components of each result are:

- Result number
- Document title (*Clicking on this will take you to the original document*)
- Abstract (*The first few hundred characters of the document*)
- Match quality graph. 84%***** ____ (*Only shown if relevance ranking was used*)
- Size (*How big is the original document*)
- Depth (*How many clicks from the top of the site*)

- Find Similar (*Find other documents similar to this one*)
- Match Info (*View the matches and other information about the document*)
- Show Parents (*List pages that link to this one*)

6.4.1 Viewing Match Info

The `Match Info` link will show you the context of your answers within the matching document. Matching words will be shown as hyperlinks. Clicking on any match term will take you to the next matching term. A summary at the top of the in-context view shows information about the document, including the time it was last modified.

6.4.2 Finding Similar Documents

The `Find Similar` link will find documents that are similar to the corresponding result. It does this by reading the original document to ascertain its main subject matter, and then conducting a relevance ranked search for those subjects.

Result documents are ordered from best to worst match. The bar graph display will indicate the overall quality of the match.

Note: The document you click on may not be ranked as the best match. This is because other documents may contain more information about the overall subject matter than the original.

6.4.3 Showing Document Parents

Often it is difficult to navigate using a search engine because there is no *back-link* present on the matching document. The `Show Parents` link solves this.

This link will show other documents that contain hyperlinks to the one you click on. In other words, it is an automated back button.